Young Scholars Journal

№ 1 – 2 2022

PREMIER
Publishing

Vienna
2022

Young Scholars Journal
Scientific journal
№ 1 – 2 2022

ISSN 2519-9331

Proofreading Kristin Theissen

Cover design Andreas Vogel

Additional design Stephan Friedman

Editorial office Premier Publishing s.r.o. Praha 8 – Karlín, Lyčkovo nám. 508/7, PSČ 18600

E-mail: pub@ppublishing.org

Homepage: ppublishing.org

Young Scholars Journal is an international, German/English/Russian language, peer-reviewed journal. It is published bimonthly with circulation of 1000 copies.

The decisive criterion for accepting a manuscript for publication is scientific quality. All research articles published in this journal have undergone a rigorous peer review. Based on initial screening by the editors, each paper is anonymized and reviewed by at least two anonymous referees. Recommending the articles for publishing, the reviewers confirm that in their opinion the submitted article contains important or new scientific results.
Premier Publishing s.r.o. is not responsible for the stylistic content of the article. The responsibility for the stylistic content lies on an author of an article.

Instructions for authors

Full instructions for manuscript preparation and submission can be found through the Premier Publishing s.r.o. home page at:
http://ppublishing.org

Material disclaimer

The opinions expressed in the conference proceedings do not necessarily reflect those of the Premier Publishing s.r.o., the editor, the editorial board, or the organization to which the authors are affiliated.

Premier Publishing s.r.o. is not responsible for the stylistic content of the article. The responsibility for the stylistic content lies on an author of an article.

Included to the open access repositories:

Typeset in Berling by Ziegler Buchdruckerei, Linz, Austria.
Printed by Premier Publishing s.r.o., Vienna, Austria on acid-free paper.

Section 1. History and archaeology

https://doi.org/10.29013/YSJ-22-1.2-3-7

Ronnie Wei,
Toronto, Canada in grade 11 with university aspirations

THE POPE AND THE CRUSADES

Abstract. While one can point to countless examples of events that altered the history of European, and to a larger extent, world history, there is no doubt that the crusades is one of those events. Lasting over a century, the conflicts between the crusader nations and the middle east solidified the Pope's legitimacy, bolstered the European nation's economy for centuries to come, and increased the Catholic influence on continental Europe extending into Asia. Using data, textbooks, and journals, this research paper examines the crusades from the perspective of the twenty-first century. The research only invites speculations for what the world would be like today without the crusades and centuries of catholic influence that followed.

Keywords: Crusades, Catholic, Europe, Pope.

Introduction

Lasting from 1095 to 1291, the Crusades remain one of the most influential events in both Christian and Islamic history. It has shaped the destiny and history of countless European and Middle Eastern nations. While its impact on humanity and the involved religions and countries is undeniable, scholars have long debated the true victors of the Crusades. However, one thing is for certain: the Pope rose to power from the Crusades. Deviating from popular interpretations and beliefs, this paper argues that, at the conclusion of the crusades, neither Islamic nor Christian states benefited more than the Pope. As a direct result of the crusades, the Pope gained wealth, authority, and legitimacy.

Origin of the Pope

The word "pope" derives from πάπας, which is Greek for father. The title was applied to all senior clergy, nobility, and bishops in early Christianity. However, the role itself did not become official until centuries later. According to the Roman Catholic Church, the first recorded Pope was Saint Peter, known as Peter the Apostle, who was one of Jesus's twelve apostles. He became Pope of Rome in AD30. His tenure ended between Ad 64 and 68. Saint Peter is widely regarded by the Catholic Church as one of the first leaders of the early church.

According to the Bible, Saint Peter was entrusted by Jesus with the "keys to heaven" (Matthew 16:19). He also oversaw the Roman Catholic Church and its bishops. Due to his close relations with Jesus, who promised him a 'special position in the Church,' Peter's authority was not

questioned. Presently, the Pope functions as the head of the entire Catholic church, having supreme powers over the future of the Church and an influence on Catholic states.

The Start of the Crusades

Fast forward to the second millenium. One hundred and fifty eight Popes later, Urban II emerged.

Becoming Pope in 1088, Urban felt that his most urgent task was to secure his position against the antipope, Clement III, and to establish his authority as legitimate Pope throughout Christendom in Europe.

The Holy Roman emperor at the time, Emperor Henry IV, quarreled with the Pope, and he subsequently faced a potential rebellion among the German nobles. At the same time, Clement III became the Italian leader of the imperialist faction opposing the Gregorian regime. After Pope Gregory VII excommunicated him, Clement was elected antipope on June 25, 1080, by Henry at Brixen. Clement III reigned in opposition to two successive popes, Gregory VII and Victor III. As an antipope, he opposed the legitimately elected bishop of Rome, endeavored to secure the papal throne, and changed the manner of choosing the pope.

Pope Urban railed against simony and other clerical abuses prevalent during the Middle Ages. Furthermore, on 27 November 1095, Pope Urban II delivered perhaps the most influential speech of the Middle Ages, calling for the start of the Crusades. He summoned all Christians in Europe to war against Muslims in the Middle East in order to reclaim the Holy Land. His cries of "Deus volt!" and "God wills it!" were met with an astonishing response, both among the military elite as well as ordinary citizens. Those who joined the crusader armies wore a cross as a symbol of the Church and loyalty to the Pope.

The First Crusades

Four armies of crusaders were formed from troops of different Western European regions, led by Raymond of Saint-Gilles, Godfrey of Bouillon, Hugh of Vermandois, and Bohemond of Taranto (with his nephew Tancred). These groups departed for Byzantium in August 1096. After some exchanges, in May 1097, the crusaders and their Byzantine allies attacked Nicea (now Iznik, Turkey), the Seljuk capital in Anatolia. The city surrendered in late June, marking the end of the First Crusade. Having achieved their goal in an unexpectedly short period of time, many crusaders departed for home. To govern the conquered territory, the crusaders who remained established four large western settlements, or Crusader states, in Jerusalem, Edessa, Antioch and Tripoli.

However, peace was short-lived. Muslim forces began gaining ground in their own holy war against the Christians. In 1144, Seljuk general Zangi, governor of Mosul, captured Edessa, leading to the loss of the northernmost Crusader state. Europe was caught by surprise with the sudden Muslim attacks on the Crusdaer states. As a result, Christian authorities in the West called for another crusade. Led by two great rulers, King Louis VII of France and King Conrad III of the Holy Roman Empire, the Second Crusade began in 1147. After they managed to assemble their armies at Jerusalem, they attacked the major Syrian city, Damascus, with an astounding army of around 50,000, the largest crusader army yet. The rulers of Damascus called on Nur al-Din, who was Zangi's successor in Mosul, for support. Together, they defeated the crusaders army, which concluded the Second Crusade. What followed were years in which the crusaders campaigned to capture Egypt. They were defeated by Nur al-Din's forces in 1169. Following his death, Saladin became his successor. In 1187, he began his major campaign against the crusader's major city, Jerusalem. His forces recaptured important cities along the way, which sparked outrage amongst the crusaders and started the Third Crusade. In September 1191, King Richard I of England's army defeated Saladin at the Battle of Arsuf, which was the only true battle of the entire Third Crusade. One year later, Richard and Saladin signed a peace treaty to reestablish the Kingdom of Jerusalem, which successfully ended the Third Crusade.

In 1204, the crusaders declared war on Constantinople, capital of the Byzantine Empire, which started the Fourth Crusade. It ended when Constantinople fell, which nearly destroyed the ancient capital.

Small Crusades

During the 13th century, numerous smaller crusades were started, with goals to combat any groups seen as enemies of the Christian church and the

Pope. For example, the Albigensian Crusade tried to eliminate the heretical Cathari and Albigensian sects of Christianity in France. The Fifth Crusade in 1216 started by Pope Innocent III saw crusaders unsuccessful attempt to secure Egypt from the hands of the Muslim. In 1229, Emperor Frederick II started the Sixth Crusade when he gained control over Jerusalem from the Muslims with a peace treaty. However, just a decade later, it was returned. Spanning from 1248 to 1254, Louis IX organized the Seventh Crusade against Egypt again, but it came to no success. In 1291, one of the only remaining Crusader cities, Acre, fell to the Muslim. This event was the final straw; this defeat marked the end of the Crusader States and the Crusades themselves.

Victory of the Pope

The Economic Gains of the Pope

Despite many historians believing the Crusades were ultimately unsuccessful, the Roman Catholic Church surprisingly gained financially despite their numerous losses. Meanwhile, the Pope gained more power than ever. One reason for this were improvements in trade and transportation throughout Europe. The Crusades were a big reason for Europe's wealth increase. The wars created a constant demand for supplies and transportation, which resulted in increased ship-building and manufacturing of various supplies. After the Crusades, there was an increased interest in European travel and learning throughout Europe. Some historians believe this may have paved the way for the Renaissance. Also, the Pope's influence over Catholic nations in Europe was further established during the Crusades. Many of the Crusader States have become world leaders today (Heston [5, 122]). Even though the papacy lost all of its territorial gains in Asia, the Crusades were still an economic and political success in Europe.

the Crusades marked the end of nearly two hundred years of clashes between the west and east. Both sides were devastated by these encounters. Equally, however, both sides benefited. The Pope and the Church received great endowments for the long crusades, which contributed to his expanded wealth. Men who had taken the Cross and were unable to go, purchased exemptions with their vows. This went to the Church and eventually to the Pope. Taxes for the Crusades were frequently collected and handled by the Church. It is not possible to give any estimate of the total amount which the Church received through the Crusades, but it was an enormous sum (Munro [4, 5]). As a result of the listed factors from the Crusades, the Pope became much more powerful and rich, especially through his control over the appointment of the officials who profited by his wealth. Additionally, Crusader States gained new ways to trade intercontinentally, which provided ways for many new tourists around the world to visit Europe, either for diplomacy or simply leisure. During the same time, between 800 and 1200, Europe began to emerge from the Dark Ages. During this time, the "Classical Era of Islam" took place. Major cities under Muslim leadership increased emphasis on science, medicine and technology. In 1312, The Council of Vienna was established as a center for Arabic, Greek, Hebrew, and Syriac studies, further solidifying European countries' international knowledge after the Crusades (Heston [5, 125]). While the development of Europe after the Crusades was prevalent, so was the Church and the Popes, who gained an enormous increase in wealth. Crusaders gave freely to the Church before marching towards the East. Crusaders also mortgaged or sold their property to ecclesiastical foundations under conditions very advantageous to the Church.

Increase in the Pope's Authority

Moreover, the Pope also made reforms to ensure more capital would end up going to the Church. The Church sought control of potential wealth through the priesthood and children of concubines or other out-of-wedlock relationships. The Church instituted celibacy vows for priests, which made it increasingly possible that familial inheritances would go to the Church. Another reform the Church instituted was prohibiting adoptions, which transferred rights of

inheritance in the case of the childless. This may account for two other offsetting institutions, namely, orphanages and homes for the elderly or abandoned. In terms of family structure, there were fewer families, with perhaps more children per family. The Catholic Church accumulated large amounts of land as a result, which was met with much envy from the royalty in France, England, and elsewhere (Heston [5, 129]).

Indeed, the Pope and Church gained economically after the Crusades. He reasserted his authority over Europe and beyond. During the Crusades, the Pope found ways to tighten his control on many monarchies in Europe and established legislation that diminished opposition powers. The Pope gave permission for non-payment of debts owed by crusaders and ordered monarchs across Europe to reinforce the rule. This encroachment upon property rights was one example that provoked less opposition because the creditors were frequently Jews. As each crusader was under the protection of the Church, the Popes interfered in case of capture of individual crusaders by their enemies and also to prevent warfare, which would have hindered men from fulfilling their vows. They censured the Church freely for this purpose; it was met with general approval. The Pope even interfered with the amusements of nobility and clergy, repeatedly forbidding tournaments and threatening to excommunicate all participants. These are just a few notable instances of the Pope adding to his power and control over those who were not members of the clergy. After a century of crusading activity, the Pope's power had enormously strengthened in Europe (Heston [5, 121]).

To grow his following and control even more territory during the Crusades, the Pope offered privileges to anyone who took the Cross. Because of the intense enthusiasm for the Crusades and also because of the weakness of most of the monarchs in Western Europe during the first half of the twelfth century, the Church, and especially the Pope, encroached upon temporal authorities. All crusaders were given the protection of the ecclesiastical courts; thus when a vassal took the Cross. he might escape to a considerable extent from the jurisdiction of his feudal lord. Moreover, crusader families and properties were taken under the protection of the Church. In this way, many cases were taken from the feudal courts. During the Crusades, outbursts of religious enthusiasm led to the Children's Crusade, which was to be a missionary movement not a military campaign. Thus, throughout the Crusades, there was a great amount of religious fervor, some real reformation in manners, and a greater interest in the Holy Land and Catholic Church. This would redound to the credit of the Pope and increase his influence and power (Munro [4, 350]). At the end of the century that saw the conclusion of the Crusades, saw the temporary union of almost all Christian lands under the authority of the Pope. This was directly due to the Crusades. The capture of Constantinople led to the establishment of a Latin patriarchate there. Furthermore, Bishops of heretical churches in Syria acknowledged the supremacy of the Latin Church, while the rulers of Armenia sought to have the title of king bestowed by the Pope and promised in return to bring the Armenian Church under the Pope. At the time, the possibility that there might be one Catholic Church takeover under the authority of the Pope was legitimate (Munro [4, 351]).

Solidification of Legitimacy for the Pope

The Pope also cemented his legitimacy at the conclusion of the Crusades. However, his cementation of legitimacy did not only come as a result of the Crusades. During the mid-eleventh century, the Church did not have the institutional means nor the "moral authority" to employ armed forces in pursuit of their crusading interests. In order for the Catholic Church to pursue the idea of the Crusades, it had to meet two conditions. First, the Church would have to be reconstituted as a legitimate war unit. This meant that it would have to be transformed into a corporate entity with the widely accepted legitimate authority to employ military forces in the pursuit of its interests. Second, the armed nobilities who made up the crusader armies would have to be reconstituted as "soldiers of Christ" (*milites Christi*) willing and able to fight on behalf of the Church and its

interests (Latham [2, 234]). Due to the establishment of these forces in the name of the Church, and with the Pope being the head of the Church, his undeniable position would only be further strengthened.

With the backing of a legitimate army fighting in honor of the Church, the Pope preached crusades against his Sicilian kingdom. This further illustrates another means by which the power of the Pope was enhanced. The Pope repeatedly preached crusades against the Church's temporal foes and offered the participants the same privileges, spiritual and temporal, which were given to those who went on expeditions against the Moslems. These holy wars were sometimes directed against monarchs and other rulers, sometimes against cities, at other times against heretics like the Albigenses, or against the heathen in the north and northeast of Europe (Munro [4, 352]). These armies played an important part in legitimizing the Pope's influence in the thirteenth century.

Conclusion

The lasting consequences extending from the Crusades would not exist without the initiation of the Pope. Likewise, it only makes sense that the Pope would be the one to reap the rewards of the Crusades. Through the Pope's increasing control on politics, it enabled him to create regulations beneficial to the Catholic Church and the Crusader States; it created the first crusader army, allowed the Church to benefit economically, expanded his following

tremendously and legitimized the Pope's position in the Church. The aftermath of the Pope's influence during the Crusades is still present today. Not only does the Pope still hold the title of being the head of the Catholic Church nearly one meilleum later, but some of the Crusader States

have developed into world leaders today. This outcome invites the question: what would be different today without the Pope and the Crusades?

References:

1. Chevedden Paul E. "The Islamic View and the Christian View of the Crusades: A New Synthesis." Wiley 93,– No. 2. 2008.– P. 181–200. Accessed: – June 30, 2021. URL: https://www.jstor.org/stable/24428429
2. Latham Andrew A. "Theorizing the Crusades: Identity, Institutions, and Religious War in Medieval Latin Christendom." Wiley, – 55.– No. 1. 2011.– P. 223–243. Accessed: – July 3, 2021. URL: https://www.jstor.org/stable/23019520
3. Baldwin M. W., Madden. Thomas F. and Dickson Gary. "Crusades". Encyclopedia Britannica,– December, – 29. 2020. URL: https://www.britannica.com/event/Crusades
4. Munro Dana C. "The Popes and the Crusades". Proceedings of the American Philosophical Society, – 55.– No. 5. 1916.– P. 348–56. Accessed: – July 21, 2021. URL: http://www.jstor.org/stable/984051
5. Heston Alan. "Crusades and Jihads: A Long-Run Economic Perspective". The Annals of the American Academy of Political and Social Science 588. 2003.– P. 112–35. Accessed: – August, 2. 2021.URL: http://www.jstor.org/stable/1049857

Section 2. Medical science

https://doi.org/10.29013/YSJ-22-1.2-8-17

Yue Wang,
Place of study: A high school junior at George School
City, country: Pennsylvania, United States

A STUDY OF US EXPENDITURES ON CANCER TREATMENT WITH DATA ANALYSIS AND MACHINE LEARNING

Abstract. Cancer is the second leading cause of death around the world, causing cancer cost to be an important social issue in the United States. News reports show that American cancer patients spent more than $21 billion on their care in 2019. (US News [2]) In this research, data analysis has been done based on the national expenditure on cancer treatment from 2010 to 2020 through the use of Python language and available third party libraries. Also, a machine learning classification model has been trained, developed and tested to help predict the cost of cancer treatment in the next few years. Among four different machine learning regression algorithms that are applied (i.e linear regression, lasso regression, random forest regression, and gradient boosting regression), gradient boosting regression is the best fit for the model, aiming to produce the most accurate prediction to inform people and government officials.

Keywords: cancer cost, correlation, machine learning, linear regression, lasso regression, random forest regression, gradient boosting regression.

Introduction

As the second leading cause of death, cancer has a high risk of taking people's lives away, which brings about questions regarding the cost of cancer treatment. A news report recently reveals that "American cancer patients spent more than $21 billion on their care in 2019." (US News [2]) This figure emphasizes how much US citizens have spent on cancer treatment and to what extent some of them must have suffered from the high cost brought by different kinds of cancer. In addition, this figure is increasing every year – $190.2 billion in 2015 and $208.9 billion in 2020, an increase of 10 percent that is due to aging and growth of the US population. (National Cancer Institute [5]) To stress the heavy cost of dealing with cancer, *U. S. Bureau of Labor Statistics* provided data contrasting between American people's average monthly income pre-tax and their average cost of cancer treatments per month. It is obvious to see from the comparison the inequality between how much people earned and how much they spent, in which income is about $3600, while cost is about $20000. (Karen Selby [3]) Aside from the out-of-pocket medical costs, expenditure spent on commuting also accounts for the large amount of spending. For example, among the approximate $21 billion of cancer cost in 2019, $16.22 billion was spent on out-of-pocket medical costs and $4.87 billion on traveling expenses. (US News [2]) The high costs of cancer treatment exert especially great pressure on

the poor, those who are uninsured or underinsured, and blue-collar workers who may lose wages as a result of health issues. (Karen Selby [3]) According to the survey, 23% of US citizens aged 19–64 were "underinsured" in 2018 since their out-of-pocket health care costs were equal to 10 percent or more of their yearly income, meaning that insurance cannot cover their expenses on cancer treatment. (Karen Selby [3]) On the other hand, blue-collar workers are facing serious situations too, not only because they are less likely to have employer-based insurance coverage than white-collar workers, but also due to the fact that their annual mean wages weigh much less than the monthly cost of some cancer drugs, causing the cost of cancer to be unaffordable for them. Under such circumstances, people are expecting ways to lower the financial burden caused by cancer treatment, which not only means to reduce cancer patients' out-of-pocket costs, but also to address the long-term financial impact (The Cost of Cancer [5]).

In this research paper, with the data collected in *Data.World* (xprizeai-health [1]) on the estimation of the national expenditures for cancer care from 2010 to 2020 under different assumptions of cancer incidence and survival trends, data analysis is going to be completed through the use of Python language and available third party libraries. Furthermore, it is possible to predict the cost of cancer treatment in the US in the next few years based on previous analysis through the use of machine learning algorithms.

Data analysis

The data used in this study is from *Data.World* (xprizeai-health [1]), which is the world's largest collaborative data community. This dataset is an estimation of the national expenditures for cancer care from 2010 to 2020 in billion dollars under different assumptions of cancer incidence and survival trends, including 1258 entries of different kinds of cancer cost. The descriptions of column features and their corresponding values are shown below in Table 1:

Table 1.– Description of column features

Type	AllSites, bladder, brain, breast, cervix, colorectal, esophagus, head_neck, kidney, leukemia, lung, lymphoma, melanoma, ovary, pancreas, prostate, stomach, uterus, other
Year	Numerical type, including 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, and 2020
Sex	Both sexes, females, males
Age	All ages
Survival	Different combinations – (incidence, survival at constant rate), (incident follows recent trend, survival constant), (survival follows recent trend, incidence constant), (incidence, survival follow recent trends)
Cost_increase	Annual cost increase, numerical type, including 0%, 2%, and 5%
Cost_total	Total costs, numerical type
Cost_initial	Initial year after diagnosis cost, numerical type
Cost_continue	Continuing phase cost, numerical type
Cost_last	Last year of life cost, numerical type

Table 2.– Below presents 5 sample rows from the dataset:

	Type	Year	Sex	Age	Survival	Cost_increase	Cost_total	Cost_initial	Cost_continue	Cost_last
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
0	AllSites	2010	Both Sexes	All ages	Incidence, Survival at constant rate	0%	124565.6	40463.5	46642.8	37459.2

1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
1	AllSites	2010	Both Sexes	All ages	Incidence follows recent trend, Survival constant	0%	122420.8	38552.7	46671.9	37196.3
2	AllSites	2010	Both Sexes	All ages	Survival follows recent trend, Incidence constant	0%	125397.7	40463.5	47136.3	37797.9
3	AllSites	2010	Both Sexes	All ages	Incidence, Survival follow recent trends	0%	123236.3	38552.7	47155.7	37527.8
4	AllSites	2010	Both Sexes	All ages	Incidence, Survival follow recent trends	2%	123236.3	38552.7	47155.7	37527.8

Table 2: *Sample data*

The next process is to learn the dataset by analyzing the data and understanding the relationships between these different columns through the use of Python Data Analysis (Pandas) and Python Data Visualization Libraries (Matplotlib and Seaborn).

First, through the use of countplot in Python, which shows the counts of observations in each categorical bin using bars (Geeksfor Geeks [4]), I make a plot about different types of cancer sites shown in Figure 1 below. In this plot, each bar represents the number of counts of different types of cancer. According to the same height of bars, we can conclude that data is very evenly distributed across different cancer sites in this dataset.

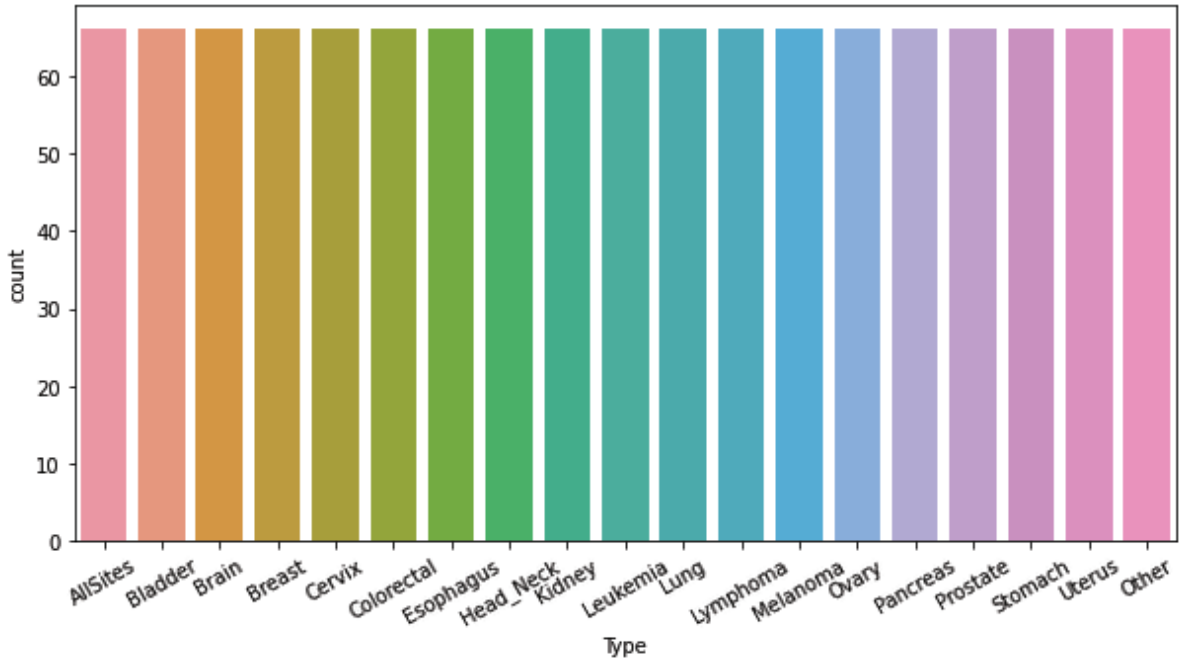


Figure 1. Plot for Type

The same measurement can be used for creating the plot of different sexes shown in Figure 2 below. In this plot, each bar represents different types of sexes – females, males, and both sexes considered as a whole. Evidently, the bar of "both sexes" is the highest, and the bar of "females" is the second highest, which is higher than the bar of "males", meaning that the population of females is larger than that of males included in this dataset.

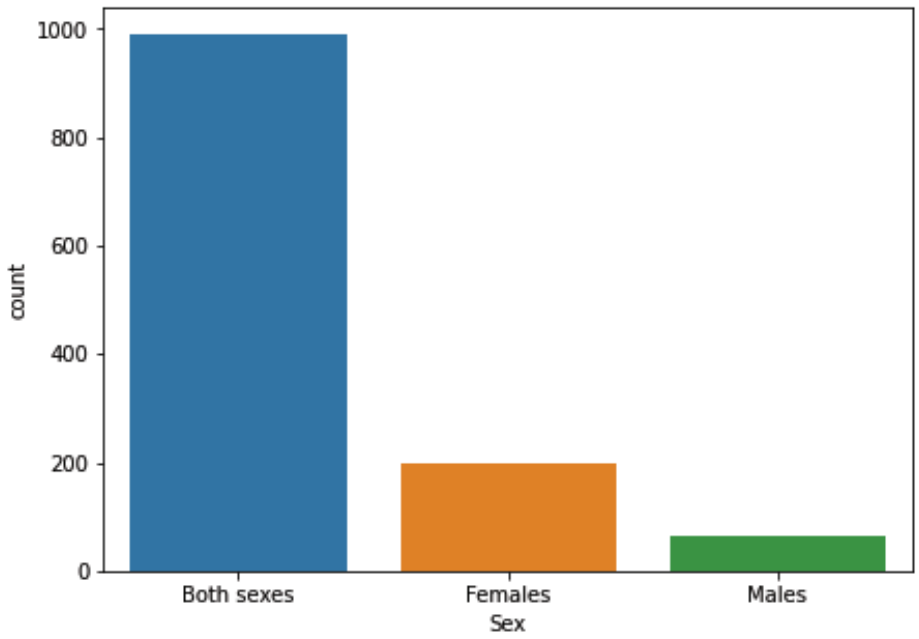


Figure 2. Plot for Sex

Also, the plot of survival and incidence was created as shown in Figure 3 below. In this plot, each bar represents different combinations of survivals – incidence and survival at constant rate, incidence follows recent trend and survival constant, survival follows recent trend and incidence constant, and incidence and survival follows recent trends. We can make the conclusion that since the first three heights are the same, the counts for them are the same, which are much lower than the fourth one.

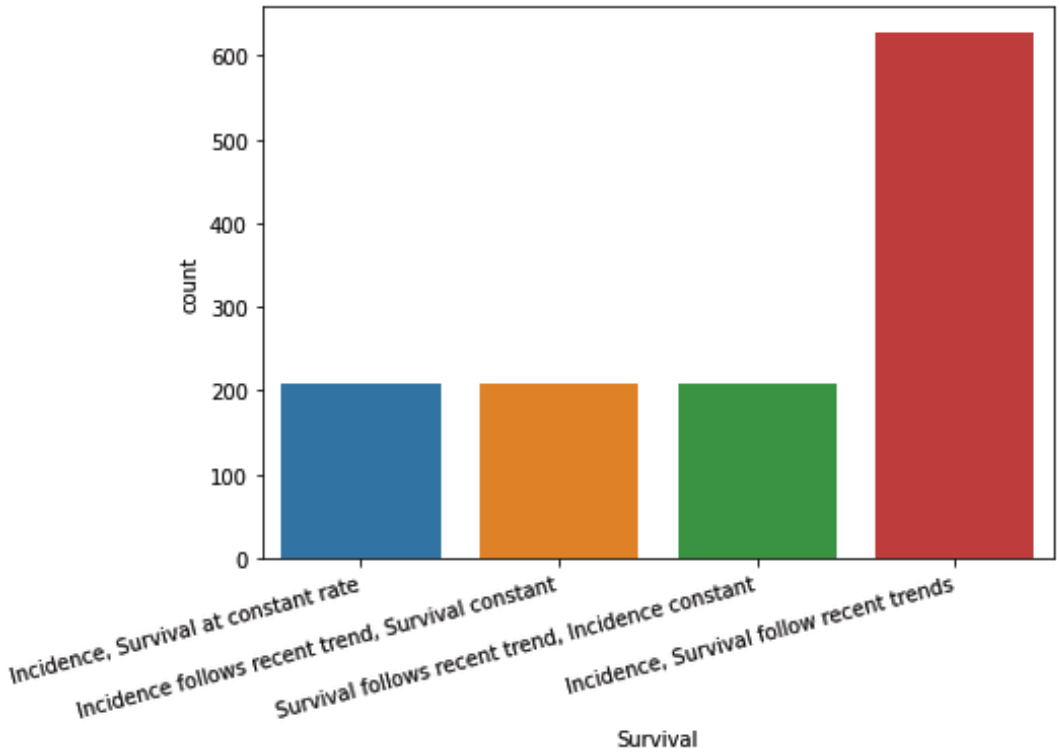


Figure 3. Plot for Survival

Next, I plot the histograms to investigate the distribution of numeric columns. As shown in Figure 4, we can draw some conclusions. For example, the data about years is evenly distributed since the number of counts for specific years is the same. For different percentages of increasing cost, the frequencies of 2% and 5% are roughly the same, which are much smaller than that of 0%. On the other hand, the rest of four distributions about costs are rightly skewed with some outliers.

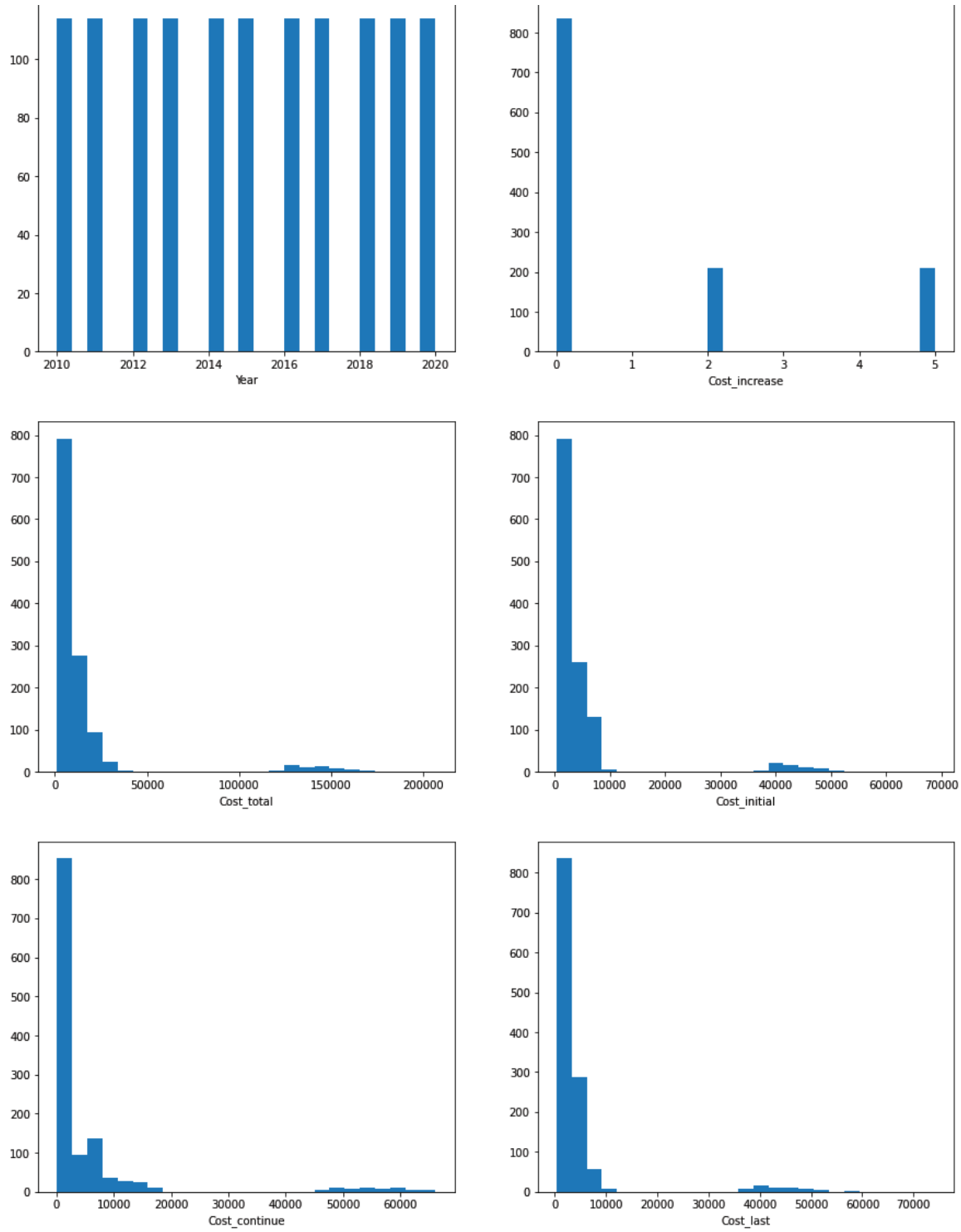


Figure 4. Distribution plots for each feature column

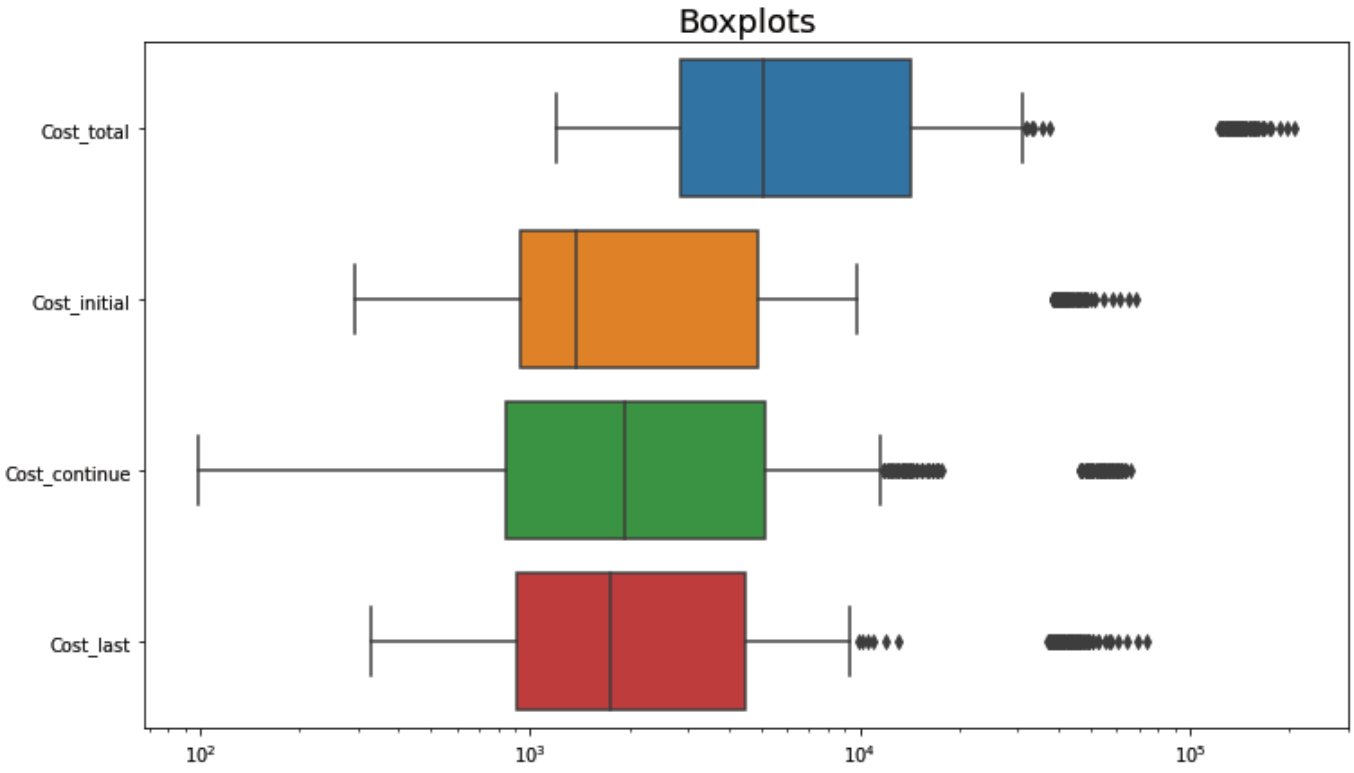


Figure 5. Boxplots for Costs

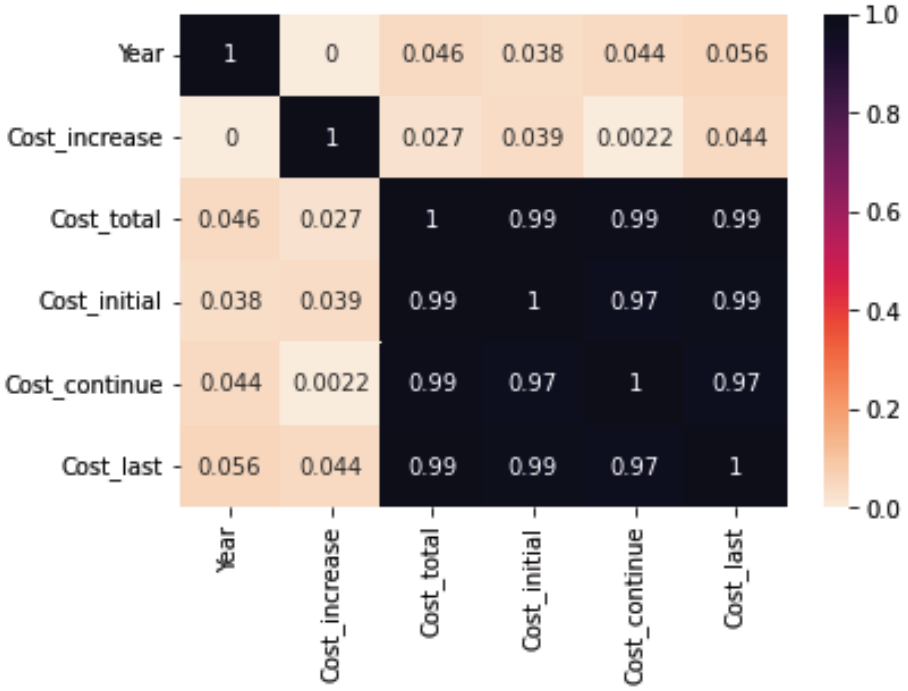


Figure 6. Correlation heatmap

To further explore these outliers, boxplot was made for specific kinds of cancer costs, in which we can have a clear idea about the values of minimum, Q_1, median, Q_3, and maximum as well as the distribution of these outliers (Figure 5).

Correlation heatmap is also produced to visualize the correlation matrix. A correlation number of 1 or close to 1 indicates that two columns are highly correlated, while a correlation number of 0 or close to 0 shows that two columns hardly correlate with each other. In accordance with the figure below, year and annual cost increase serves as two factors that are not much influencing. On the contrary, the total costs, the costs of initial year after diagnosis, continuing phase costs, and costs of last year of life have strong correlation with each other, which is rather understandable since these four factors represent various costs during different stages in the cancer treatment process.

Machine learning

This section describes the approach to develop the machine learning regression model for prediction of cancer costs.

Preprocessing

Preprocessing is an important procedure before feeding the data into the model, which is also an integral step since the output of the model can be directly influenced by the quality of the data.

The specific step of preprocessing is below:

1) One hot encoding the sex column. According to Table 1, some column features are categorical variables, and some are numerical variables. Sex is a one of these categorical variables, which needs to be converted to numerical variables for the machine learning algorithm to understand. In this way, one hot encoding serves as a good way to prepare data for an algorithm. (Dinesh Yadav [6]) For the sex column, after the conversion, all entries of "Both_sex" become "1" and those of "Male" and "Female" become "0".

2) One hot encoding the survival column. According to Table 1, survival is also a categorical variable. Through one hot encoding, it is separated into four different columns – "Incidence, survival at constant rate" to "Both_constant", "Incidence follows recent trend, survival constant" to "ConstSurv_TrendIncid", "Survival follows recent trend, incidence constant" to "ConstIncid_TrendSurv", and "Incidence, survival follows recent trends" to "Both_trend." Whether being "1" or "0" depends on what kind of survival features a specific individual has.

3) Converting the "Cost_increase" column. Since the original "Cost_increase" column is object type due to the percentage sign, which is difficult to be used in an algorithm, 0%, 2%, and 5% are changed into 0,1, and 2 respectively.

4) Frequency encoding the type column. According to table 1, there are 10 different types of cancers exhibiting in the type column, which result in too many unique values. Therefore, I attempt to group those values by frequency since I don't want 10 new columns. By mapping values to dataframe and dropping the original column, a new column "Type_freq_encode" is created.

After all four steps, the 10 new sample rows from the dataset are below shown in Table 3:

Table 3. – New Sample Data

	Year	Cost_increase	Cost_total	Cost_initial	Cost_continue	Cost_last	Both_sex	Sex_F	Sex_M	ConstSurc_ TrendIncid	Both_constant	Both_trend	ConstIncid_ TrendSurv	Type_freq_en- code
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*
0	2010	0	124565.6	40463.5	46642.8	37459.2	1	0	0	0	1	0	0	0.052632
1	2010	0	122420.8	38552.7	46671.9	37196.3	1	0	0	1	0	0	0	0.052632
2	2010	0	125397.7	40463.5	47136.3	37797.9	1	0	0	0	0	0	1	0.052632
3	2010	0	123236.3	38552.7	47155.7	37527.8	1	0	0	0	0	1	0	0.052632

1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*
4	2010	1	123236.3	38552.7	47155.7	37527.8	1	0	0	0	0	1	0	0.052632
5	2010	2	123236.3	38552.7	47155.7	37527.8	1	0	0	0	0	1	0	0.052632
6	2010	0	3980.7	978.7	1895.8	1106.3	1	0	0	0	1	0	0	0.052632
7	2010	0	3885.2	923.3	1872.3	1089.7	1	0	0	1	0	0	0	0.052632
8	2010	0	3987.7	978.7	1900.2	1108.8	1	0	0	0	0	0	1	0.052632
9	2010	0	3891.9	923.3	1876.5	1092.2	1	0	0	0	0	1	0	0.052632

We also obtain the frequency for different types of cancer, which are identical for every cancer, so we can drop this column.

Regression models

For this research, I apply four different machine learning regression algorithms, including linear regression, lasso regression, random forest regression, and gradient boosting regression. During each run, I first apply the train set to train the model, then use the model on test data to make predictions which is to test the model's performance.

First, linear regression is the base model that can be used to perform basic regression tasks. Mean absolute error is a typical metric used to evaluate a regression model, which with respect to a test set is the mean of the absolute values of the individual prediction errors on over all instances in the test set. Based on our data, the mean absolute error is about 0.0366985. Besides, r square is a necessary statistical measure of how close the data are to the regression line (Minitab, 2013), which leads to the result of about 0.999, meaning that the model fits the data well.

According to the previous correlation heatmap shown in **Figure 6**, four cost factors are highly correlated with each other. As a result, we can apply lasso regression to address the multicollinearity, which is a good solution to reduce the magnitude of the coefficients of the model while keeping other features the same (Andrea Perlato [10]). Based on our data, the mean absolute error is about 0.55354.

A random forest is a supervised machine learning algorithm used for classification and regression that is constructed from decision tree algorithms. (Afroz Chakure [6]) It is a bagging technique, which operates by constructing a multitude of decision trees at training time and outputting the mode of classes or mean prediction of trees. The figure below shows how a random forest works.

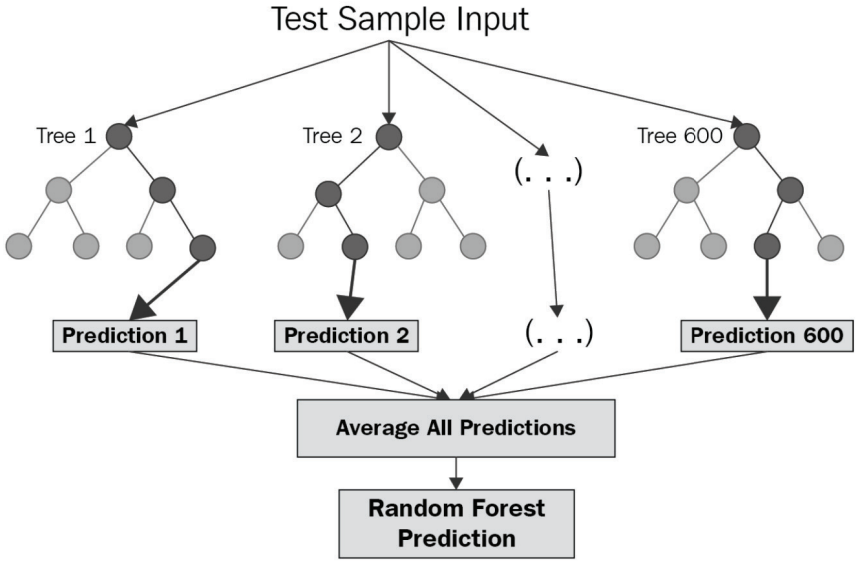


Figure 7. Random Forest Structure

According to the random forest regression, the R square value is about 0.99567.

Also, gradient boosting is also a machine learning algorithm used for classification and regression that is constructed from decision tree algorithms. Different from random forest, gradient boosting is a boosting technique, which works by building weaker prediction models sequentially where each model tries to predict the error left by the previous one. According to the gradient boosting regression, the R square value is about 0.99896.

Since the target variables are very skewed as the values of R square are rather close to 1, we should address the skewness before applying any regression model. Therefore, log transformation is an appropriate way that can be used to remove skewness from the predictor. (Dario [9]) By log transforming y, the skew coefficient changed from 3.8372 to 0.9349.

Besides, to address the multicollinearity problem, since Cost_continue, Cost_last, and Cost_initial all are very correlated with each other as the correlation coefficient is about 0.99, we can drop them and only keep Cost_initial in consideration.

By using linear regression and lasso regression model again for new transformed data, the values of R square are about 0.53246 and 0.52546 respectively. We also compare the R squares inferred from both random forest regression and gradient boosting regression, which are about 0.934734 and 0.93939285 respectively. Therefore, gradient boosting regression is the most appropriate model since its R square value approaches 1 the most, meaning that the model fits the data best.

Conclusion

In this research, four different machine learning regression algorithms have been applied to develop and train the cancer cost prediction model, which are linear regression, lasso regression, random forest regression, and gradient boosting regression. Since several factors are highly correlated with each other, we use log transformation to reduce the skewness. Based on our data, gradient boosting regression is the best machine learning algorithm since its R square is higher than that of other three algorithms, which is about 0.93939285, meaning that it fits the data well. Therefore, we can use gradient boosting regression to predict the cost of cancer treatment in the United States in the future. And we also hope to improve the dataset in order to produce the model with higher accuracy.

References:

1. Watson IBM. "Expenditures for Cancer Care – Dataset by Xprize Ai-Health." Data.world, 19 July 2017. URL: https://data.world/xprizeai-health/expenditures-for-cancer-care/workspace/project-summary?agentid=xprizeai-health&datasetid=expenditures-for-cancer-care
2. "Cancer Costs U. S. Patients $21 Billion a Year." US News. URL: https://www.usnews.com/news/health-news/articles/2021–10–26/cancer-costs-us-patients-21-billion-a-year
3. Selby Karen. "Americans Can't Keep Up with the High Cost of Cancer Treatment." Mesothelioma Center – Vital Services for Cancer Patients & Families, 20 Aug. 2021. URL: https://www.asbestos.com/featured-stories/high-cost-of-cancer-treatment
4. "Financial Burden of Cancer Care." Financial Burden of Cancer Care, 20 July 2021. URL: https://progressreport.cancer.gov/after/economic_burden.
5. "The American Cancer Society Cancer Action NetworkSM (ACS CAN). Is Making Cancer-and the Affordability of Cancer Care-a Top Priority for Public Officials and Candidates at the Federal, State and Local Levels". The Costs of Cancer, Oct. 2020. URL: https://www.fightcancer.org/sites/default/files/National%20Documents/Costs-of-Cancer-2020–10222020.pdf/ Accessed: 16 Dec. 2021.

6. Yadav D. Categorical encoding using label-encoding and one-hot-encoder. Medium. (2019, December 9). Retrieved January – 22, 2022. URL: from https://towardsdatascience.com/categorical-encoding-using-label-encoding-and-one-hot-encoder-911ef77fb5bd
7. Editor M. B. (n.d.). Regression analysis: How do I interpret R-squared and assess the goodness-of-fit? Minitab Blog. Retrieved January 22, 2022. URL: from https://blog.minitab.com/en/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit
8. Chakure A. Random Forest and its implementation. (2020, November 6). Medium. Retrieved January 22, 2022. URL: from https://medium.com/swlh/random-forest-and-its-implementation-71824ced454f
9. Radečić D. Top 3 methods for handling skewed data. (2020, January 4). Medium. Retrieved January 22, 2022. URL: from https://towardsdatascience.com/top-3-methods-for-handling-skewed-data-1334e0debf45
10. Parleto A. Deal multicollinearity with lasso regression. (2020). Andrea Perlato. Retrieved January 22, 2022. URL: from https://www.andreaperlato.com/mlpost/deal-multicollinearity-with-lasso-regression

Section 3. Psychology

https://doi.org/10.29013/YSJ-22-1.2-18-22

Sikun Gan,
Coventry Christian School
High School Student

BIG DATA EMOTION CLASSIFICATION

Abstract. Affecting our mind in terms of decision-making, influencing our moods and behaviors, emotion makes up a major part of our daily lives. People's physical and mental health and work status can be adversely affected by persistent negative emotions, but positive emotions can enhance subjective well-being and promote physical and mental health. I then formulated a question that can positive/negative emotion be automatically classified using a model, i.e, does the sentence contains enough information for the computer to make a sentiment judgment. With a massive amount of text data, I here build up an automatic emotion classification model that could read and distinguish sentences with negative emotions from sentences with positive emotions. Specifically, I studied the penalized logistic regression model with Stanford movie review data as the input. The AUC metric is used for model evaluation and outputted a promising out of sample score of 0.96.

Keywords: Logistic regression, big data, sentiment classification.

1. Introduction

Emotion makes up a major part of our daily lives. It affects our mind in terms of decision-making, influences our moods and behaviors, and so forth. To name a few, there are investment strategies developed based upon the investor's emotion gathered from stock comments. There are automatic question and answering systems that feed users with different answers based upon the emotion in the online chat window. In general, understanding one's emotions can extrapolate his behaviors, actions, and mental health condition. With the recent advancement of technology, most people can access the Internet and social media. The majority can share their emotions via many different platforms, such as Twitter, Instagram, and Tik Tok. Therefore, by utilizing user-generated content in the correct manner, we will be able to gauge people's mental health. It could be possible to predict mental health levels and depression by mining the content from social media platforms. Depression is a serious medical condition that hampers the ability to perform normal daily tasks such as working, studying, eating, sleeping and having fun. Thanks to the rapid growth of computer utility, we can develop automatic tools for classifying and identifying emotions behind the posts today by incorporating databases and algorithms.

2. Stanford Movie Review Dataset

To study the emotion classification, I adopt the Stanford Movie Review Dataset which contains 25,000 reviews for popular movies. The data is in text format with the following print out examples:

```
label: 1 review: For a movie that gets no respect there sure are a lot of mem
label: 1 review: Bizarre horror movie filled with famous faces but stolen by
label: 1 review: A solid, if unremarkable film. Matthau, as Einstein, was won
```

The dataset has been used and cited many times in natural language processing field.

3. Data Preparation

The data is in the text format which requires me to conduct data cleaning first. One way of extracting the emotion information from the text is to count how many times does the positive emotion related words show up in the text. One way to count the occurrence of the keywords in a sentence is to use the CountVectorizer function from python package sklearn.

However, such a method is not ideal because some of the keywords with high frequency are meaningless in the sense that it appears in 99% of the text. For example, the words "the" and "a" are commonly used in all contexts of English language. To better extract the useful keywords summarizing the sentence emotion, I adopt the term frequency-inverse document frequency, also known as TF-IDF. Specifically, Term frequency is how many times does a word appears in the comment, which is defined:

$$TF = \frac{time\ of\ occurence}{total\ number\ of\ words\ in\ the\ article}$$

To calculate IDF, we need a corpus that contains every possible word. The formula should be

$$IDF = log\left(\frac{the\ number\ of\ document}{the\ number\ of\ document\ contains\ a\ word + 1}\right)$$

Finally, we can get TF-IDF by mulitiplying TF and IDF, TF * IDF. By adopting TF-IDF, we can summarize the information within the movie comments with a list of important keywords. The Sklearn package also provide a function Tfidf Vectorizer, which facilitate my research for TF-IDF computation.

4. Model

With the TF-IDF numerical data matrix as the input, I adopt the logistic regression model for emotion classification. Because the data contains too many parameters, I applied L1 penalization.

4.1 Model setup

The logistic regression models the probability of the labeled data. It has wide and successful application in statistics. The model starts by modeling the probability of the comment to be positive with the following equation:

$$Log\frac{p}{1-p} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

where:

- p is the probability for the comment to be positive;
- $X_1, X_2, \ldots X_p$ are p keywords appearing in the columns of the TF-IDF matrix.

The model normalizes the RHS equation to domain $[0,1]$ but one can always obtain the probability for a comment to be positive with the following equation:

$$p = \frac{exp\left(\beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots + \beta_p X_p\right)}{1 + exp\left(\beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots + \beta_p X_p\right)}$$

4.2 Binomial Probability

To connect the data to the probability model, the logistic regression assumes to use $Y = 1$ to indicate the emotion is positive and to use $Y = 0$ to indicate the emotion is negative. For example, suppose we observe a sentence/comment with positive emotion, the probability for that sentence being positive is p and the probability for that sentence being negative is 1-p, which can be rewritten in the following form:

$$P(Y) = p^Y (1-p)^{1-Y}$$

If we observe more than one sentence, assuming each of the sentence being independent, we have their joint probability being modeled by the product of each individual probability:

$$P(Y_1, Y_2, Y_n, \ldots, Y_n) = \pi_{i=1}^{n} p_i^{Y_i} (1-p_i)^{1-Y_i}$$

Intuitively, we know that depending on the text used in the review sentence, the probability for its emotion to be positive should be different from

sentence to sentence. This coincides with the product of the binomial probability above because the model assumes the probability for the i-th sentence being positive to be p_i. From the previous section, we know that p_i can be modeled with the following equation:

$$p_i = \frac{exp\left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} \ldots + \beta_p X_{ip}\right)}{1 + exp\left(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} \ldots + \beta_p X_{ip}\right)}$$

4.3 Model solution

The above model is now defined on a set of parameters β. The model solution is found by maximizing the probability $\pi_{i=1}^{n} p_i^{Y_i} \left(1 - p_i\right)^{1-Y_i}$ due to the assumption that the probability model we propose should best explain the data by letting the observed data have the highest probability. To facilitate the maximization, the model is usually solved by taking the log operation, after which we could apply the chain rule from calculus to solve for the maximum point:

$$argmax_\beta \frac{1}{N} \sum_{i=1}^{n} y_i log\left(p_i\right) + \left(1 - y_i\right) log\left(1 - p_i\right)$$

For my research, I used python Sklearn package to solve for $(\beta_0^*, \beta_1^*, \ldots, \beta_p^*)$ that maximize $\pi_{i=1}^{n} p_i^{Y_i} \left(1 - p_i\right)^{1-Y_i}$

4.4 Penalized Logistic Regression

Ideally, we should know that some of the $(\beta_0^*, \beta_1^*, \ldots, \beta_p^*)$ to be zero because the TF-IDF matrix contains a lot of 0s. To force some of the parameters to be 0, the model can be extended to penalized logistic regression, which minimizing the following equation instead of solely maximizing the joint probability.

$$argmin_\beta -\frac{1}{N} \sum_{i=1}^{n} y_i log\left(p_i\right) - \left(1 - y_i\right) log\left(1 - p_i\right) + \\ + \lambda\left(\left|\beta_0\right| + \left|\beta_1\right| + \ldots \left|\beta_p\right|\right)$$

The first term in red is essentially the negative of the log probability, minimizing of which is equivalent to maximizing the negative of itself. The second summation is a penalization term where if some of the β_i s are not zero, then it will bring up the optimization function value. Essentially, the model is looking for a solution of $(\beta_0^*, \beta_1^*, \ldots, \beta_p^*)$ that maximizes the binomial probability while using the fewest number of βs.

5. Result Analysis

5.1 Train-Test Dataset Split

After we find our model solution $(\beta_0^*, \beta_1^*, \ldots, \beta_p^*)$, given a new sentence X_{new}, we will be able to find the positive sentiment probability by

$$p_i = \frac{ex\left(\beta_0^* + \beta_1^* X_{i1}^{new} + \ldots + \beta_p^* X_{ip}^{new}\right)}{1 + ex\left(\beta_0^* + \beta_1^* X_{i1}^{new} + \ldots + \beta_p^* X_{ip}^{new}\right)}$$

We could simply adopt 0.5 as the cutoff value to make emotion classification decisions. Specifically, if the probability is above 0.5, we set the sentence emotion to be positive, otherwise, we set the sentence emotion to be negative. To better evaluate my model, I used 75% of the data for finding my parameters $(\beta_0^*, \beta_1^*, \ldots, \beta_p^*)$ and pretend the rest 25% of the data as new observations for model evaluation. Statistically speaking, the 75% data used is called the training set and the rest 25% of the data is called testing set.

5.2 ROC Curve

To evaluate the model performance, we could simply use the accuracy score

$$Accuarcy = \frac{TP + TN}{TP + TN + FP + FN}$$

Where TP = True positive; FP = False positive; TN = True negative; FN = False negative.

However, such a evaluation might not be accurate if the data has 98% of the review comments to be positive, in which case, one naïve model of rating all the comments to be positive would give us 98% accuracy. To comprehensively evaluate the model performance, I study the False Positive Rate(FPR) and True Positive Rate(TPR):

$$TPR = \frac{TP}{TP + FN}, FPR = \frac{FP}{FP + TN}$$

Under the FPR evaluation, if a model simply rates all the comment to be positive, it will have a low FPR.

Another issue with the model evaluation is that adopting 0.5 as the cutoff value makes intuitive sense but might not be the best approach. In reality, we

might want to be conversant by setting the cutoff value to be 0.6, or even 0.9. Each choice of the cutoff value can thus give us different FPR and TPR. To comprehensively measure the model performance with different cutoff value, I study the ROC curve defined below.

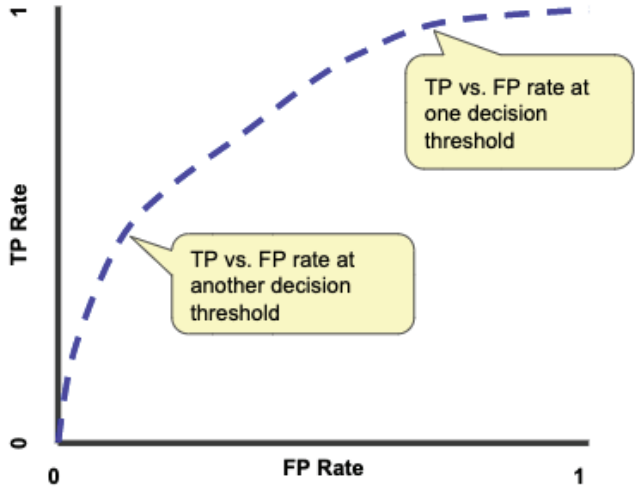


Figure 1.

Where the x-axis is the FPR and the y-axis is the TPR. For each cutoff value, we can obtain a pair of (TPR, FPR) and label it on the above coordinate. If we keep change the cutoff value from 0 to 1, we get many dots and if we connect those dot, it gives us the ROC curve. A good model thus should have the curve close to the top-left corner because it indicates a lower FPR and a high TPR. We know that the Area of the coordinate is 1 and the higher the area under the ROC curve, the better the model is. Consequently, the Area Under the Curve (AUC) is a natural judgement on the classification model performance.

I finally evaluate my model with the 25% testing set on the ROC curve and obtain the following ROC plot, which has AUC score of 0.96.

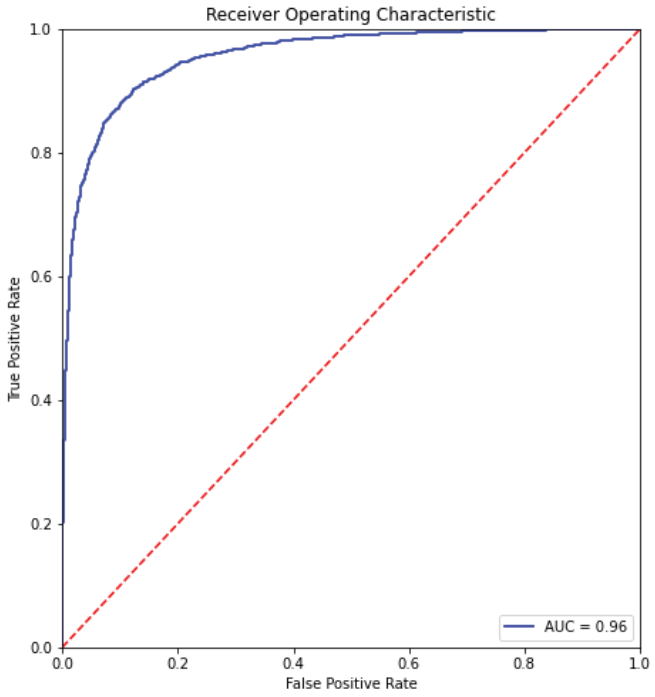


Figure 2.

6. Conclusion

With the TF-IDF metric to summarize the sentence information and with penalized logistic regression to model the sentiment probability, I build up a probability model for sentiment classification. To take the sample imbalance into consideration, I also researched the model evaluation by studying the confusion matrix and ROC/AUC metrics. The out of sample AUC score, 0.96, indicates that my penalized logistic regression can distinguish the negative movie comments from positive comments with high accuracy. With the preliminary exploration of emotion classification, I can further research to create a model that is able to classify different types of emotions beyond mere two polarities, thus can estimate people's depression levels with more confidence and maintain their well-being.

References:

1. All of Statistics: A Concise Course in Statistical Inference, Larry A. Wasserman. 2004.
2. Linear Regression Using R: An Introduction to Data Modeling, David J. Lilja. 2016.
3. Sentiment Analysis. URL: https://ai.stanford.edu/~amaas/data/sentiment
4. Learning Word Vectors for Sentiment Analysis, Proceedings of the 49 th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011.

5. Scikit-Learn: Machine Learning in Python – Scikit-Learn 1.0 Documentation. URL: https://scikit-learn.org/stable
6. Classification: ROC Curve and AUC | Machine Learning Crash Course. Google Developers. URL: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

https://doi.org/10.29013/YSJ-22-1.2-23-29

Qinglan Luo,
11th-grade student at Rutgers Prep

AUTISM AMONG CHILDREN IN2019 NATIONAL SURVEY OF CHILDREN'S HEALTH

Abstract. Autism is a broad range of complex developmental disabilities in social communication and interaction coincided with restricted, repetitive behaviors. According to research from the National Survey of Children's Health (NSCH), approximately 28.5% of children aged from 0 to 17 have autism, and the percentage has been increasing in recent years. To control the increasing trend of autism, this study aims to examine the predictors of autism and build a predictive model for autism using the logistic regression model.

We used the 2019 NSCH dataset in this report. After feature normalization, we built a logistic regression model to predict whether a child is likely to develop autism. The predictive model is further validated by an overall evaluation of the model and has achieved an AUROC score of 0.68. By investigating the correlation among variables and the logistic regression coefficients, we found the dependent variable is most positively correlated with the financial situation of the household and most negatively correlated with the gender of the child. The results imply that older children are more likely to develop autism ($p = 0.0086$, OR = 1.093), and children in the family whose income is not able to cover basic living are more likely to develop autism ($p < 0.001$, OR = 1.77). Besides, children do not have a low birth weight ($p = 0.033$, OR = 0.47) and female children are less likely to develop autism ($p < 0.001$, OR = 0.82).

Keywords: autism, financial situation, predictive model, machine learning, logistic regression.

1. Introduction

"It is like getting an internet error when you are playing a game. The little avatar on the screen is disconnected from the outside world. He was just trapped there, unable to move or get out." This is a heartbreakingly true account given by a child's mother with autism. Before the age of two and a half, Liu had always been a bright and lively child. While his growth trajectory was no different from that of a normal child, Liu's symptoms were suddenly detected by his parents – he became muted just overnight. Then, he was diagnosed with autism. The abrupt diagnosis shocked Liu and his family, signaling a cascade of questions and inviting an ever-evolving emotional burden with it. Luckily for Liu, he was not the first autistic child to be clinically assessed in the United States. Nowadays, there have been more analogous cases in the United States. In order to control the growth of such tragedies, scientists have done plenty of research on the causes of ASD. It has been clear now what genetic factors play an essential role in leading to this disease. On the other hand, the environmental factors are still implicitly demonstrated.

Autism, also known as the autism spectrum disorder (ASD), is still an undeveloped area of understanding. It is a broad range of complex developmental disabilities in social communication and interaction coincided with restricted, repetitive behaviors. According to DSM-5, social deficits are defined as "deficits in social-emotional reciprocity, deficits in nonverbal communicative behaviors used for social interaction, and deficits in developing, maintaining, and understand relationships". (CDC, 2020) Repetitive behaviors include "stereotyped

or repetitive motor movements, use of objects, or speech, insistence on sameness, inflexible adherence to routines, or ritualized patterns of verbal or nonverbal behavior, highly restricted, fixated interests that are abnormal in intensity or focus, hyper- or hyporeactivity to sensory input or unusual interest in sensory aspects of the environment." (CDC, 2020) According to research from the National Survey of Children's Health (NSCH), there were approximately 28.5% of children aged from 0 to 17 having autism (CDC, 2021), increasing by 9.7 in percentage compared to the number of autisms in 2012 (188 per 1000 children). Indeed, there has been a consistent increase in the number of diagnosed autisms across various data sources. Although people are still not sure to what extent the changes in the clinical definitions of ASD and more people being consciously involved in ASD diagnosis has stimulated the growth, it is dangerous to assume that the real number of ASD has been stable and in control in the recent years.

The statistics have shown that this disorder transcends race, gender, and SES and challenges teenagers in different degrees. According to DSM-5, there are three levels of autism. Level 1 refers to any diagnosed children having trouble initiating social interactions and requiring support; level 2 includes children whose social interactions are limited to narrow special interests and have frequent restricted or repetitive behaviors; level 3 refers to children with severe deficits in verbal and nonverbal social communication skills, requiring substantial support from others (CDC, 2020).

As the number of people with autism rises each year, more families are suffering from the disease, and many talented children cannot utilize their gifts because of autism. Although we are still unclear about the treatment yet, we can go into the causative factors of this disorder and effectively prevent its increase after fully understanding both the genetic and environmental factors. This study serves this end by aiming to examine the predictors of autism and building a predictive model for autism using the logistic regression model.

2. Data and Methods:

2.1 Data

This report uses data from the National Survey of Children's Health (NSCH) in 2019, which is a population-based survey established by the Health Resources and Services Administration (HRSA) Maternal and Child Health Bureau (MCHB) to monitor the prevalence of the children health condition in the United States and to evaluate their access to quality health care (NSCH – Questionnaires 2019). The whole survey mainly encapsulates family composition, race/ethnicity, income, type of health insurance, and a variety of other important demographic and health status characteristics related questions. The data is collected by telephone surveys to random households across the United States. The 2019 NSCH dataset is used in this report. Before the data-cleaning process, the NHIS dataset has 67,625 valid observations.

The table below shows all the variables that have been chosen in this report to examine the relationship between independent variables and the dependent variable:

Table 1.– Variables used for analysis

Item Code	Question
1	*2*
HHCOUNT	How many people are living or staying at this address?
A1_SEX	What is your sex?
A1_BORN	Where were you born?
A1_GRADE	What is the highest grade or level of school you have completed?
A1_MARITAL	What is your marital status?
A1_AGE	What is your age?

1	*2*
A1_PHYSHEALTH	In general, how is your physical health?
A1_MENTHEALTH	In general, how is your mental or emotional health?
SC_AGE_YEARS	Is this child 3 years old or older?
SC_RACE_R	What is this child's race/ethnicity?
SC_HISPANIC_R	Is this child Hispanic?
BIRTHWT_L	Is this child born with low birth weight?
ACE1	How often has it been very hard to cover the basics, like food or housing, on your family's income?
ACE3	To the best of your knowledge, has this child EVER experienced: parent or guardian divorced?
ACE4	To the best of your knowledge, has this child EVER experienced: parent or guardian died?
ACE5	To the best of your knowledge, has this child EVER experienced: parent or guardian served time in jail?
ACE6	To the best of your knowledge, has this child EVER experienced: saw or heard parents or adults slap, hit, kick punch one another in the home?
ACE7	To the best of your knowledge, has this child EVER experienced: was a victim of violence or witnessed violence in his or her neighborhood?
ACE8	To the best of your knowledge, has this child EVER experienced: lived with anyone who was mentally ill, suicidal, or severely depressed?
ACE9	To the best of your knowledge, has this child EVER experienced: lived with anyone who had a problem with alcohol or drugs?
ACE10	To the best of your knowledge, has this child EVER experienced: treated or judged unfairly because of his or her race or ethnic group?
AUTISMMED	Is this child CURRENTLY taking medication for Autism, ASD, Asperger's Disorder or PDD?

This report uses the variable "AUTISMMED" as the dependent variable. Responses to the question "AUTISMMED" is dichotomous, meaning that the respondents either answer "yes", indicating that the child needs treatment for autism, or "no", indicating that the child does not need such treatment or does not have such behavior.

2.2 Statistical Models

2.2.1 Pre-processing

The data set is pre-processed in this step to improve both the training speed and accuracy. Since there is inevitably missing data, imputation is required to better analyze and extrapolate the missing data. As most machine learning algorithms are not able to deal with missing values, we replaced the missing values with the mean value of the entire feature column (mean value imputation). Some machine learning algorithms, such as artificial neural networks, require a specific technique called feature scaling which transforms different features into comparable scales for better training speed and accuracy. In this report, we will use the min-max scalar for this purpose. For each feature, its minimum and maximum value are first computed as X_{min} and X_{max}. Then each data point X with respect to that feature is replaced by X_{sc} calculated as:

$$X_{sc} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Using this formula, X_{sc} is the ultimate value that is going to be analyzed in this report.

2.2.2 Logistic Regression Model

A logistic regression model refers to a model that is used to predict the probability of an incidence to happen. The probability varies from 0 to 1, with zero indicating not likely to happen and one indicating very likely to happen. Instead of a linear relationship, the logistic regression model fits an "S" shape which can be expressed using the formula below:

$$\ln\left(\frac{y}{y\text{-}1}\right) = a_0 + a_1x_1 + a_2x_2 + L + a_nx_n$$

In the above equation, a_0 is the intercept, xn represents the independent variables, and a_1 to a_n, are their corresponding coefficients (weights). In this report, our goal is to find the coefficients $(a_0, \ldots, a_n)$ minimizing the sum of squared errors (SSE) so that our predicted values will deviate the least from the real values.

2.3 Model Validation

Consider a two-class prediction problem, where the outcomes are labeled either as positive or negative. There are four possible outcomes from a binary classifier. If the outcome from a prediction is positive and the actual value is also positive, then it is called a true positive (TP); however, if the actual value is negative then it is said to be a false positive (FP). Conversely, a true negative (TN) has occurred when both the prediction outcome and the actual value are negative, and a false negative (FN) is when the prediction outcome is negative while the actual value is positive. In this way, the true positive rate (TPR) can be calculated as follows:

$$TPR = \frac{TP}{TP + FN}$$

And the false positive rate (FPR) can be calculated as:

$$FPR = \frac{FP}{TN + FP}$$

A receiver operating characteristic curve, or ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied (Google, 2020). The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The best possible prediction method would yield a point in the upper left corner of the ROC space. A random guess would give a point along a diagonal line from the left bottom to the top right corners. Points above the diagonal represent better than random classification results, while points below the line represent worse than random results. In general, ROC analysis is one tool to select possibly optimal models and to discard suboptimal ones independently from the class distribution. Sometimes, it might be hard to identify which algorithm performs better by directly looking at ROC curves. Area Under Curve (AUC) overcomes this drawback by finding the area under the ROC curve, making it easier to find the optimal model.

3. Results

3.1 Chorogram

A chorogram is a graphical representation of the cells of a matrix of correlations. The idea is to display the pattern of correlations in terms of their signs and magnitudes by using visual thinning and correlation-based variable ordering. Moreover, the cells of the matrix can be shaded or colored to show the correlation value. The positive correlations are shown in blue, while the negative correlations are shown in red; the darker the hue, the greater the magnitude of the correlation.

According to the chorogram above, children's chance for developing autism has the strongest positive correlation with the variable "ACE1", whether the household is able to cover basics on family's income, and has the strongest negative relationship with "SC_SEX", which is the gender of the child.

3.2 Logistic Regression Results

The results of logistic regression analysis of children ever need medical treatment for emotional and behavioral disorder are listed in the figure below.

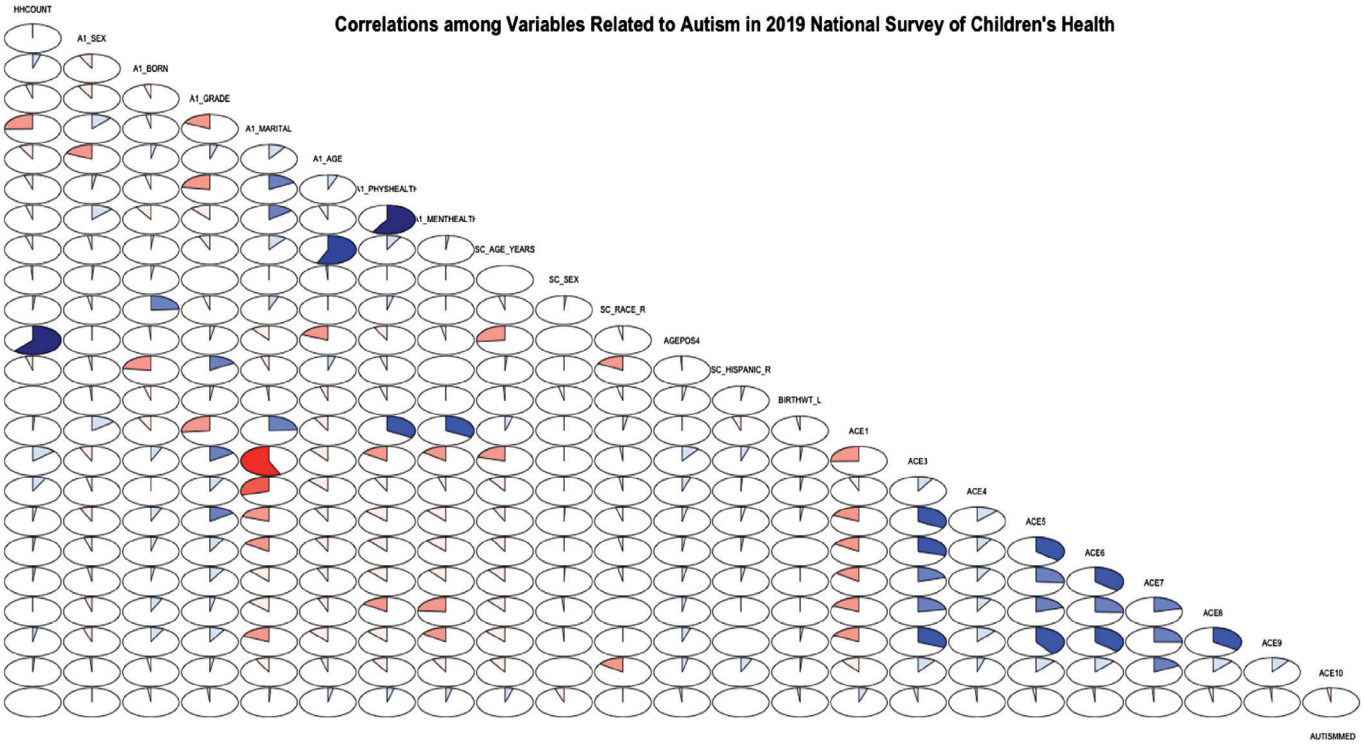


Figure 1. Correlation among variables

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -2.07213    2.66074  -0.779 0.436109
HHCOUNT        0.08447    0.13388   0.631 0.528043
A1_SEX        -0.19398    0.28884  -0.672 0.501855
A1_BORN       -0.67205    0.54575  -1.231 0.218170
A1_GRADE       0.01166    0.07533   0.155 0.877036
A1_MARITAL    -0.15882    0.12512  -1.269 0.204328
A1_AGE         0.01486    0.01638   0.907 0.364441
A1_PHYSHEALTH  0.31368    0.17159   1.828 0.067535 .
A1_MENTHEALTH  0.01666    0.17218   0.097 0.922911
SC_AGE_YEARS   0.08895    0.03384   2.628 0.008579 **
SC_SEX        -1.29141    0.31505  -4.099 4.15e-05 ***
SC_RACE_R     -0.08852    0.08788  -1.007 0.313782
AGEPOS4       -0.00577    0.18131  -0.032 0.974614
SC_HISPANIC_R -0.38219    0.38751  -0.986 0.324007
BIRTHWT_L     -0.76067    0.35645  -2.134 0.032839 *
ACE1           0.57006    0.14735   3.869 0.000109 ***
ACE3          -0.21313    0.33652  -0.633 0.526521
ACE4          -0.88675    0.49241  -1.801 0.071730 .
ACE5           0.17044    0.52198   0.327 0.744031
ACE6           0.01877    0.55622   0.034 0.973085
ACE7           0.67314    0.65041   1.035 0.300701
ACE8           0.13867    0.41484   0.334 0.738178
ACE9          -0.06939    0.43282  -0.160 0.872623
ACE10         -0.66089    0.51839  -1.275 0.202349
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
                [,1]
(Intercept)   0.1259168
HHCOUNT       1.0881461
A1_SEX        0.8236755
A1_BORN       0.5106616
A1_GRADE      1.0117235
A1_MARITAL    0.8531473
A1_AGE        1.0149668
A1_PHYSHEALTH 1.3684558
A1_MENTHEALTH 1.0168010
SC_AGE_YEARS  1.0930289
SC_SEX        0.2748817
SC_RACE_R     0.9152824
AGEPOS4       0.9942471
SC_HISPANIC_R 0.6823669
BIRTHWT_L     0.4673508
ACE1          1.7683658
ACE3          0.8080552
ACE4          0.4119904
ACE5          1.1858233
ACE6          1.0189436
ACE7          1.9603741
ACE8          1.1487411
ACE9          0.9329604
ACE10         0.5163909
```

Figure 2. Logistic regression results (left), odds ratio for each variable (right)

From the logistic regression results, it is not hard to find that, taking a 90% confidence level, children's age, sex, physical health, birth weight, whether the parents or guardians are dead, and whether the household is able to cover daily basic life are all significant predictors of the dependent variable. More specifically, according to the logistic regression coefficients, children's sex, birth weight, and whether the parents or guardians are dead are significant negative predictors, while physical health, age, and whether a family's income can cover basic living are significant positive predictors. The results of the logistic regression model corroborate with the findings from the correlogram shown above.

In addition, combing with the odds ratio table, we could identify those older children are more likely to develop autism ($p = 0.0086$, $OR = 1.093$), children with a worse physical health condition are more likely to develop autism ($p = 0.067$, $OR = 1.093$), and children in the family whose income is not able to cover basic living are more likely to develop autism ($p < 0.001$, $OR = 1.77$).

Besides, children who do not have a low birth weight ($p = 0.033$, $OR = 0.47$), whose parents or guardians are not dead are less like to develop autism ($p = 0.071$, $OR = 0.41$), and female children are less likely to develop autism ($p < 0.001$, $OR = 0.82$).

3.3 Model Validation

The figure below displays the ROC curve for the logistic regression model with an AUROC score 0.68. It can be concluded that the model has achieved a relatively good performance much better than randomly guessing.

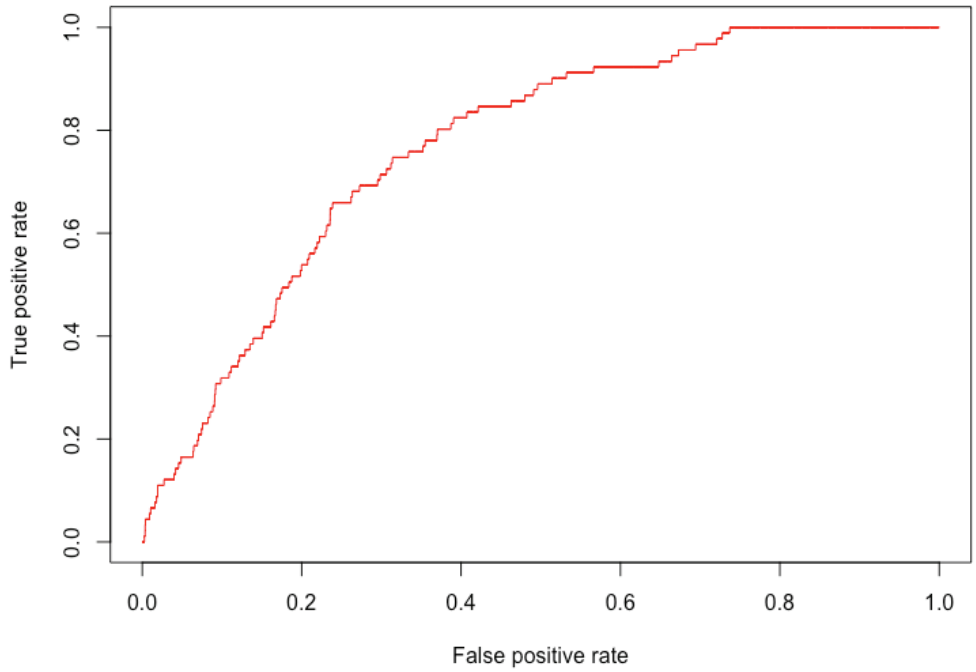


Figure 3. ROC curve

4. Discussions

The intention of this study is to build a predictive model with the best performance and to investigate the factors most related to children's chance of developing autism. One logistic regression model was built and has achieved good performance much better than randomly guessing. Also, from the logistic regression results, we are able to ascertain that the child's age, sex, physical health, birth weight, family income, and negative childhood experiences are all significant predictors of the dependent variable, which corroborates with the findings from the correlogram. Combining both results, we can see that in order to assess the child's chance of developing autism, it will be most effective to look at factors such as the child's negative childhood experience and family social-economic status.

One limitation of the study is that data entries with missing values are imputed with the mean value of the entire feature column. This is a timesaving but defective approach. Depending on the number of such data entries, it is possible that we might introduce a new bias into the dataset. For future studies, we may use more advanced techniques such as k-nearest neighbors (kNN) imputation, which replaces missing values with the mean of k (a parameter selected by the user) nearest neighbors of that sample. This technique requires more efforts but can generally achieve better performance and may help create a more accurate model.

References:

1. Centers for Disease Control and Prevention. (2020, June 29). Diagnostic criteria. Centers for Disease Control and Prevention. Retrieved February 21, 2022. From URL: https://www.cdc.gov/ncbddd/autism/hcp-dsm.html
2. Centers for Disease Control and Prevention. (2021, December 2). Autism data visualization tool. Centers for Disease Control and Prevention. Retrieved February 21, 2022. From URL: https://www.cdc.gov/ncbddd/autism/data/index.html#data
3. Google. (2020 Aug. 11). Classification: ROC curve and Auc | machine Learning crash course. Google. URL: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc.
4. National Survey of Children's Health – Data Resource Center for Child and Adolescent Health, 2019. URL: https://www.childhealthdata.org/learn-about-the-nsch/NSCH

Section 4. Sociology

https://doi.org/10.29013/YSJ-22-1.2-30-45

Jiaqi Wu,
Shanghai Pinghe School, China

STUDY OF THE ETHICAL ISSUES IN TIKTOK

Abstract. In recent years, TikTok, a globally popular video-sharing smartphone application (app), has gained a reputation for capturing the zeitgeist of culture and redefining the way people use social media. This research paper aims at figuring out the ethical problem in Tiktok. After briefly describing the meaning of algorithms in life and some potential problems of using algorithms, this paper focuses on the social software TikTok, a short video application that has become very popular and controversial in recent years. This research has also used the TikTok dataset and modeling to support the spread of inappropriate text and videos. Through the analysis of the TikTok dataset and some visual data processing methods, how the recommendation system works is better displayed in this paper. By analyzing the data, the paper also makes it clear why the recommendation system brings the threat of spreading inappropriate information, both in text and videos. Additionally, the research outlines to a certain degree some challenges when solving the algorithm bias in TikTok through some examples. Furthermore, this paper analyzes the measures taken by the government and TikTok to face these problems. It emphasized that moral problems with Douyin are inevitable, but in addition to social platforms and the government, this research has concluded that there are better ways for TikTok to alleviate the problem to some extent. This paper recommends further research into more social media apps besides TikTok and further analysis of the potential problems that appeared in these apps or some common problems both share.

Keywords: Tiktok; Algorithm; Recommendation system.

1. Introduction

1.1 Algorithm

On the fast track of development, our society is moving forward by a leap. Technology does more than facilitating communication and ending the days when people used to meet for a chat. Artificial intelligence also equips people with the ability to replace a large proportion of complex and mundane work. Artificial intelligence or algorithm is a set of instructions that take an input, A, and provide an output, B, that changes the data in some processes [1]. It has widespread use on the daily basis and is still becoming more and more ubiquitous without awareness. In a golden age of data, for a myriad of jobs that people did on their own before, colossal datasets and algorithms can almost completely replace human labor. These sets of instructions for solving problems or completing tasks play an indispensable role in our everyday life: from the Internet to medical care, from education to transportations, algorithms are helping us in far more fields than we can imagine, such as calculating the most efficient routes of delivery, predicting the most profitable

marketing strategy, increase observability of specific behaviors and making a sensible public health policy.

Here is the way how the algorithm can be applied in the dataset. Sifting through masses of data, algorithms provide computers with a successive and precise guide to do specific tasks. Utilization of this can be followed back to one of the famous persons in computational hypothesis. In 1952, Alan Turing, a mathematician, published a set of equations that tried to explain the patterns people see in nature, such as the dappled stripes of zebra, the whorled leaves on a plant stem, and the complex tucking and folding that turns a ball of cells into an organism [2]. Nowadays, simply put, algorithms are also about a set of equations to solve specific problems. The equations work via input and output by applying each step of the algorithm to the data to generate the result we needed which is the output. In recent years, increasing amounts of available data have enabled AI algorithms to be utilized in many fields.

In our daily life, file compression is a process that algorithm runs on data in which the size of a file, as well as the storage, is reduced by re-encoding the data. After downloading a.zip file, operating systems can extract the contents of this.zip file like a normal folder. Rather than extract everything manually just like a decade ago, the operating systems work in the background and dive into the.zip files in a few seconds. This is because compression algorithms can optimize specifically for different files and make them into a usable form. Compression algorithms are a kind of operating system that turns the input,.zip files, into the output, usable files. Long series of bytes that are repeatedly used is found by the algorithms, and it replaces them with a single byte or as short series of bytes as possible that rarely occurs in the file [3]. Therefore, this reduces the amount of memory required to store files [4]. Just like compression algorithms, many other kinds of algorithms are playing important roles on the daily basis and providing people with a richer trove of valuable information. These algorithms are far more complicated than the equation via input and output, but the essential behind algorithms is to simulate the data processing and decision-making processes of the human brain.

1.2 Concerns of algorithms

These complex algorithms are used in many other aspects that we do not aware when using them, but is affecting us every day, such as business and marketing. "We're at the beginning of a golden age of AI. Recent advancements have already led to inventions that previously lived in the realm of science fiction – and we've only scratched the surface of what's possible." Jeff Bezos, Amazon CEO said. Nowadays, almost all the businesses are utilizing and taking advantages of AI. The improvements in algorithms make it possible to automate decision-making based on colossal database. For example, price discrimination algorithm is a practice used by businesses to charge a different price for the same product to different consumers [5]. However, increasing capabilities of price discrimination algorithms is debatable. For businesses, they implement price discrimination using algorithm by profiling consumers using cookies which is commonly used in ecommerce applications [6]. This information can be used for either consumer segmentation with different pricing displayed or personalization, such as product/service recommendation, targeted advertisement, targeted coupons with a flat price, and rank in algorithmic search which influences click-through rates [7]. These can all be approaches to implement price discriminate using algorithms. In this aspect, the outcome of algorithm is negative for consumers. For people who are charged with a higher price, it is somehow inequitable. Additionally, their personal information is gleaned by the business, which will cause some network security problems.

Concerns about algorithm is also present in other fields. With big data, a series of algorithms have significantly improved their usability on Internet and social media. Issues about using algorithm on the Internet and social media bring AI against human ethics. The questions being raised about algorithms are not about algorithms per se, but about how algorithm is used and how data is used [8].

This problem is "the social dilemma", meaning social media applications or websites manipulate people by using algorithms. Nowadays, technology companies offer a lot of free products, Facebook, WeChat, Tiktok, Twitter, for which people don't pay to use. This is attributed to that the advertisers are people who pay. The advertisers are the customers of the social applications and websites, and the users and their personal information are the goods to be sold. For example, Google's system scans the emails being sent and received by personal account and gain the content of these emails. By using information gleaned from a user's email combined with data from their Google profile, Google display what it considers are relevant advertisements [9]. Because users will only see the advertisements they are interested in, they are likely to click on it and it is harder to determine whether the messages presented by advertising are truthful or not. More often than not, by exaggerating some attributes of the product and downplaying the product's drawbacks to people who are interested in it. This also gives companies an opportunity to encourage endless purchase requests.

Surveillance capitalists sell certainty to companies that want to know with certainty what people do and what people will do. In 1986, 1% of the world's information was digitised. In 2013, it was 98% [10] and this is where the certainty comes from. By using digitized information, algorithm determine the things you receive in those websites and applications. Engineers at companies like Facebook and Google spend huge chunk of time to create new algorithms and tweaking old ones [11]. It is a matter of fact that today's internet is ruled by algorithms to some extent. Social media is no longer a bicycle waiting to be ridden, or a car to be driven. It is self-learning and self-analyzing. To some extents, artificial intelligence has gradually evolved into a tool for soliciting, luring, and even manipulating and profiting from human beings.

1.3 Background information of TikTok

Artificial intelligence is also commonly used in other social media websites and applications, including the popular video-sharing application–TikTok. TikTok is a globally popular smartphone application of sharing and creating short videos. It is owned by Byte Dance Ltd., a privately held company headquartered in Beijing, China and owned many other popular applications in China [12]. In recent years, it has gained a reputation as capturing the zeitgeist of culture and redefining the way people use social media. Initially launch in China by Byte Dance, the app become well-known around the world after renaming to TikTok. In November 2017, the previous rival of TikTok, Musical.ly, was acquired by the company for $1 billion, and the user accounts of both are consolidated [13]. Now, it has 689 million users internationally every month [14] and has downloads estimated at between 850 million and 987 million, excluding Chinese third-party downloads, in 2020 [15]. Allowing users to create, watch, and share short entertaining videos set to various music and sound effects, TikTok is notable for its high level of engagement. The personalized feeds of videos and the addictive quality further drives TikTok's popularity. By collaborating on content with other users, having online live shows, and easily adding effects like filters, stickers, and lip sync sound effects, TikTok attracts mounting number of creators and becomes the second most downloaded iPhone app in 2020 [16].

Since the app is free, it seems mysterious how the application makes money. The truth is that the main value of TikTok is the users or the gargantuan user base. With more than 2 billion users, the advertisements run by TikTok are the most obvious way it makes money. Different types of advertisements, including In-Feed ads, Brand Takeovers, and Branded Hashtag Challenges, allow TikTok to earn a large chunk of money from people or businesses who want to display their advertisements or new hashtags. When we are scrolling through the For You page, receiving an advertising videos, and seeing the newly shown up hashtags, branded custom stickers, or augmented-reality filters and lenses, TikTok is earning money and selling the products–the users. Additionally, the in-app purchase is another income

for TikTok. Using money to purchase virtual coins, people over eighteen years old are allowed to use this as in-app currency and send virtual gifts to others.

2. Ethical problem in TikTok

2.1 Newly Appeared Ethical Problem

TikTok's appeal relies on addictive quality and high levels of engagement. After opening this application, the For You page is what the users see first. Equipped with a powerful recommendation system, TikTok can automatically provide videos curated to people's interests. In the For You page, the short-form video uniquely tailored for every user is played one-by-one instantaneously as the application is being opened.

In fact, the For You page relies on algorithms to provide a myriad of videos which is different for every one of the 689 million monthly active users worldwide [17], and the success of TikTok also has a close relationship with this powerful intelligent algorithm behind it. People were speculating about this complex and mysterious For You page and how recommendations are delivered to their feed, attributed to that TikTok had been secretive about the platform's algorithms in the past. TikTok finally revealed how the For You Page recommendations are generated several years ago. However, at the same time when people know more about the algorithms. There are newly appeared ethic problems in TikTok. The data from the users is the main basis of this application, but the way of using data and feature videos makes people concern. As how videos get featured is almost only based on the For You algorithm, but lack manual review, the videos that is ethically wrong or the videos with misleading information will easily get spread. Under this circumstance, TikTok becomes a great place for malicious people to hide nude images and objectionable contents. Its lax security and control have allowed it to become a place filled with pedophiles, profanity, crime, violence and extremism and will lead to many fatal consequences on not only TikTok users, but also all people online [18].

This paper focuses on TikTok's video recommendation system and the ethical problem arising from it. By using the TikTok video dataset, the paper further analyzes why the ethical problem is caused by the recommendation algorithm. The TikTok dataset is processed by using Apache Spark to formulate a series of queries, and some analytics are extracted based on the data contained within this dataset. Furthermore, this paper analyzes the solutions to those ethical problems in terms of the algorithm bias. This research can be useful for people to have a deeper understanding of the TikTok's video recommendation system and figure out the way of improving this technology.

2.2 Technical development of video recommendation system

As mentioned before, not like other application, the addictive quality led by algorithm in recommendation system is the kernel of TikTok's success. Bytedance is focusing on the artificial intelligence in various products. An AI Lab is established in 2016, focusing on the development of innovative technologies that serve the purposes of ByteDance's content platforms [19]. In TikTok, it is also this powerful algorithm technology that leads to its popularity. TikTok is an algorithm-driven, content-oriented product. Besides the basic, industrialized recommendation engine need a robust backend and architecture design for integration and huge database to support the algorithm [20].

2.2.1 How TikTok recommends videos

According to TikTok: "The system recommends content by ranking videos based on a combination of factors – starting from interests you express as a new user and adjusting for things you indicate you're not interested in, too" [21]. From hashtags used to user interactions and your settings– they can all influence the algorithm and the videos recommended to you. With this recommendation system, the application has the capability to eliminate almost all the initiative searching of its users. It enables users to easily get the videos they like and enables videos to pass to only people who will like them. TikTok revolutionize the way for people to search. It greatly alleviates the effort people pay to obtain useful and interested

information from massive amounts of other information. With thoroughly understanding of the recommendation algorithm principles, we can better figure out the ethical problems in TikTok and the solutions. The technical development of video recommendation system in TikTok is explained by how specific users and videos are matched to each other.

In the blog posted by TikTok on June 18, 2020, "How TikTok Recommends Videos #ForYou." According to this blog, a number of factors are used in the recommendation system, some are not just considered equally important but individually weighted by TikTok's For You algorithm. The specific factors, according to TikTok, are listed below [21]:

1. User interactions such as the videos you like or share, accounts you follow, comments you post, and content you create.

2. Video information, which might include details like captions, sounds, and hashtags.

3. Device and account settings like your language preference, country setting, and device type. These factors are included to make sure the system is optimized for performance, but they receive lower weight in the recommendation system relative to other data points we measure since users don't actively express these as preferences.

According to TikTok,

All these factors are processed by our recommendation system and weighted based on their value to a user. A strong indicator of interest, such as whether a user finish watching a longer video from beginning to end, would receive greater weight than a weak indicator, such as whether the video's viewer and creator are both in the same country. Videos are then ranked to determine the likelihood of a user's interest in a piece of content and delivered to each unique For You feed [21].

This blog, to some extents, provides transparency to the public and let people know more about the recommender system. Using the factors listed above, TikTok recommend the content people love. If one clicks on a singing video, TikTok will customize according to this video and the mechanism will analyze further to trace users' future behaviors and offer precise recommendations. Without any initiative selection or search, the user passively accepts the system's personalized recommendation content [19].

2.2.2 How videos get featured

From the aspect of how videos get featured, the platform's algorithm can be explained in another way. First, the algorithm needs draw user's persona and understands the users well, even better than themselves do. By sorting the user's information and classifying the user according to their interest characteristics, the algorithm can deliver the videos accurately according to users' persona and the classification, which is User Clustering formed by the algorithm. Second, once a video is uploaded, TikTok will automatically extract features of videos in the context of recommender systems. It first exploits critical engineered features that would be derived and feed-in into the recommendation system [20]. After this, by showing it to a small number of users who are classified to be more likely to engage with it. This process of choosing a small subset of users is based on the users' persona and the likelihood of a user's interest in this piece of content. If these users watch the video in full, share it, give a like or follow the creator, which means they respond favorably, the video will be shown to more people that is also considered to have the similar interests. That same process then repeats itself in just a few seconds, and if this positive feedback loop happens continuously, the video can go viral in TikTok [22]. This process is done almost completely by the system's algorithm.

3. Analysis of the ethical problem using TikTok video dataset

3.1 Ethical problems caused by powerful algorithm in TikTok

"A Global Interest Discovery Recommendation Method and Device" is a patent ByteDance applied for how the company systematically recommends contents to users who are interested in them. As mentioned before, the powerful algorithm adopted by the recommendation system help to build a hi-

erarchical interest label set according to the topic of the content, and calculate the correlation degree of each interest label in the set in order to customize videos to the users [19]. Therefore, with this powerful algorithm, TikTok successfully attract many users and creators.

Mr. Han (translated by Erik Butler) wrote in his book: "Under the de-medialization trend in the digital media era, users are no longer just passive receivers and consumers of information, but also active senders and producers" [23]. He also said that the shitstorm represents an authentic phenomenon of digital communication [23].

People now finds TikTok not only a place to receive information, but also a platform for them to enjoy the right of expression. Even ordinary people can speak out, show their life, use this free, easy-to-use application to record everything. Therefore, with more senders and producer, TikTok is now covering all area of interests with its miscellaneous content materials. When people want to see current news, they open TikTok rather than newspaper; when they want to learn special skills, such as cooking, they open TikTok rather than books; when they want to have rest, they open TikTok, too, to watch mini self-made sitcoms, rather than open the television.

With such a huge content material on TikTok platform accurately match the corresponding audience, interesting things can quickly reach to the audience and can be spread in just a few minutes. Different from the connection system in real life between circles acquaintances, friends, and families, information in TikTok is spread without this limitation, without the narrow circle in people's real life. People have the opportunities to get access to any information they want, and, for the same, information can be spread to anyone who may be interested in it. However, TikTok not only creates the platform that dilutes the significance of social connection, but also makes it a way for malicious people to take advantage. Although there is the sophisticated and logical algorithm, which is the foundation of recommendation system, and the model used to check the content of videos, there must be unethical things being spread. TikTok embrace all the interests of users and video creators, so people are likely to be interested in and misled by fake information, objectionable content, and inappropriate material.

Therefore, not only efficient extraction of the information but also a high-quality review system is a necessary requirement. However, the review of manual editors cannot be sufficient for the rapidly changing content and a myriad of new information. With insufficient manual review, the videos that is ethically wrong or the videos with misleading information will easily get spread, even in a few seconds. It is difficult for TikTok to maintain a healthy community and eco-system with simply today's algorithm. User-generated images and videos can harm the platform and offense the users.

For example, a disturbing video of suicide was reportedly live-streamed on TikTok. The video first showed up on other social media platforms and is re-uploaded and went viral in TikTok quickly. It was not long before people starting to post the graphic video of a man live-streaming his suicide hidden in innocent contents to shock other users. Malicious people are now hiding the video in other videos to make it look like just a normal video or anything else [24]. They share it in comments, hide it in other contents, and edit it into the innocent dog and cat videos. Since this suicide video is causing negative repercussions on millions of users, the application is flooded with warnings about this viral trend, with many people warning others to exit any video pop up showing a man with long hair and a beard. Some also suggested to stay off TikTok to avoid seeing these disturbing and dangerous contents.

Theo Bertram, the European director of public policy of TikTok responded to this and said: "Through our investigations, we learned that groups operating on the dark web made plans to raid social media platforms including TikTok, in order to spread the video across the internet. What we saw

was a group of users who were repeatedly attempting to upload the video to our platform" [25].

Just like this suicide video, many other videos with inappropriate content have spread on TikTok once. Although some are removed quickly, there are many cases that the content causes a negative impact and even breaks the law. This "spreadability" is both the key draw and an ethical problem of TikTok. By facilitating the share of a new video based on the users who previously watched it and respond favorably, it is impossible for platforms to moderate the video as soon as it is posted and stop the spread in time. Users will continue to post disturbing and inappropriate content, including inappropriate language (discriminatory, racist, insults, swear words…), objectionable content, and explicit or suggestive content. There are many other cases that this kind of inappropriate material is hidden in normal contents and becomes viral in TikTok.

3.2 TikTok data analysis

3.2.1 TikTok video dataset

Analysis of the TikTok video dataset can help to get a deeper understanding of the recommendation system and extract useful and unknown patterns, relationships, and information in this recommendation algorithm. The dataset [26] used in this paper includes 165111 comments in TikTok collected from 2020.04 to 2020.10. The information includes in the dataset is all the 165111 comment ID, comment text, video ID, create time, like count, status, author unique ID, author nickname, whether the author is private, author language, author signature, author custom verification, author UID, author second UID, author avatar thumb, author region, author ins id, author youtube channel title, author youtube channel id, and author Twitter id. This TikTok dataset is worldwide data, but most in the US. The aim of using the data analysis is to show the basic rule in recommendation system. Therefore, it can support that hiding the inappropriate material into popular hashtags and terms can make it spread quickly.

3.2.2 Methodology

The data processing method is the use of Apache Spark in Python. Apache Spark is an open-source, distributed processing system that utilizes in-memory caching, and optimized query execution for fast analytic queries against data of any size [27]. It provides development APIs Python [27]. Using Apache Spark and Python, an open-source, distributed processing system used for big data workloads, some analytics about Tik Tok videos and comments are extracted by formulating a series of queries based on the data contained within the dataset.

3.2.3 Data processing

In the dataset, the text of comment and the like counts are collected and the top 5 most adopted terms by each author are depicted using the TikTok dataset and are shown in table 1.

Table 1. – Top 5 most adopted terms by each author (only showing top 10 rows)

author.unique_id	TOP5WORDS
the_stomping_lan_chop	['Hahahahahahahaha!!!!!']
jxst_vxbing	['🤣🤣🤣']
mahakhamidd	['lying', 'tho', 'not', 'he']
noneyabuisness1994	['he', 'a', 'Tesla', 'Can', 'get']
taylorjwoodward	['but', 'where', 'his', 'is', 'Tesla']
thomasmalcolm04	['@willyg04']
jillpickle31	['a', 'could', 'How', 'keep', 'you']
salma.lazraq	['not', 'him', 'getting', 'or', 'a']
someonekami17	['😂😂😂']
zacharywiseman21	['have', 'ate!!', 'David', 'for', 'a']

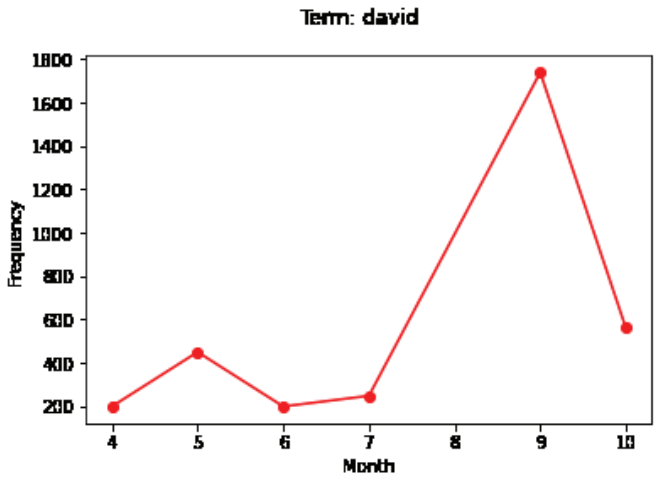
The create time of the comments and the text are collected. The 15 most frequently used terms in each of the daily time slots (i.e., Morning, Afternoon, Evening, Night) is also shown in table 2.

Table 2 – The 5 most frequently used terms in each of the daily time slots.

Time slot	Top 15 words
Evening	['a','the','he','get','you','him','I','David','tesla','Tesla','this','is','to','that','can']
Morning	['a', 'the', 'Tesla', 'him', 'tesla', 'he', 'get', 'that', 'i', 'man', 'I', 'is', 'this', 'you', 'can']
Afternoon	['a', 'him', 'get', 'this', 'I', 'the', '@daviddobrik', 'is', 'you', 'to', 'buy', 'on', 'Tesla', 'cod', 'tesla']
Night	['the', 'a', 'he', 'this', 'Tesla', 'to', 'him', 'tesla', 'is', 'I', 'that', 'Give', 'get', 'i', 'my']

Using this dataset, the text of the comments and videos are depicted. The TagCloud of terms used in TikTok is shown in the picture below. By using this visual representation of text data, keyword metadata (tags) can be depicted and can be visualized free form text. The most frequently used terms in TikTok is David, S, Tesla, Vote.etc.

Figure 1. Tag Cloud of terms



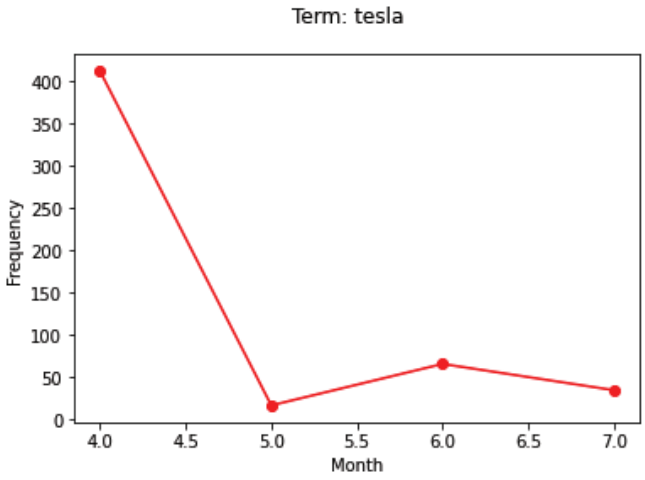


Figure 2. Use of the word "Tesla" Figure 3 Use of the word "David" over time

The text and create time are collected. Given as input a term, the graph of its use over time is shown. Figure 2 shown below is the use of the word "David" over time. There is a significant increase in the use of this word from July to September with the highest frequency reaching 1800. Apart from this period, the use of this word has a frequency of about 400. Figure 3 shown below is the use of the word "Tesla" over time. There is a decrease in the use of this word from April to May. The highest frequency in this graph is in April with a frequency of 400. Apart from this period, the use of this word has a frequency of about 50.

Given a time slot as input, the Tag Cloud of the terms used is shown. The graph shown below is the Tag Cloud of the terms used at night. The most frequently used terms at night is David, S, and Tesla.

Figure 4. Tag Cloud the terms used at night

3.2.4 Findings and conclusions

A. inappropriate text content

According to the data processing above, it can be predicted that malicious people can use the most adopted terms by each author and the most frequently used terms in a particular of the daily time slots to make their video or comment become viral quickly. By hiding the inappropriate content into the most frequently used terms in TikTok can help to hide inappropriate things and serve an illegal purpose. For example, according to the Tag Cloud, David and Tesla are the two most frequently used terms from 2020.04 to 2020.10. The term "David" become popular in August and September. This means that hiding the inappropriate content in these terms can make the videos likely go viral. Additionally, with precise calculation and prediction of what term will be used most frequently, taking advantage of the recommendation system is a great tool for illegal purposes.

Furthermore, according to the analysis of TikTok dataset, the recommendation system has a clear and predictable tendency. A particular word, even a word that is not very common in daily life, such as the term shown in the Tag Cloud, can be used much more frequently than others. When a literally number of users respond favorably to the video or comment with a particular term, the content will be shown to more people. With positive feedback loop happening enough times, the content can be spread quickly. If the content is innocent and common, the interesting content material on the platform accurately match the corresponding audience and quickly reach to the audience within a few minutes. If the term is not "David", "Tesla", "Vote", but other words or contents that may be objectionable or used to suggest inappropriate content, it can also spread very quickly since TikTok encourages creativity. Therefore, due to the nature of its algorithm, inappropriate contents are easier to go viral on TikTok [28].

For example, there are many trends in TikTok with some of the terms become very popular in a particular period of time. However, some TikTok trends have been shown to be detrimental, and even life-threatening, for young people in particular. TikTok' s users are contained by many young children or adolescents. When it comes to malicious content, they are more naïve and gullible, which means they are likely to trust the misleading information [18]. According to Statista, 25% users in TikTok are at the age of 10–19, and 21.4% users are at the age of 20–29 [29]. The high proportion of adolescents and young people in TikTok means that the dangerous terms and trends in TikTok are likely to cause negative repercussion. Young people and adolescents during puberty are often easily influenced by online information and are prone to try out what they see around them, especially when some terms or contents are very popular in Tik Tok trends. Benadryl trend can be a case in point to illustrate this. The Benadryl trend was once popular in TikTok. After some users respond favorably, the recommendation system shows it to more and more users. However, the review system does not detect the potential problem in the Benadryl trend. Benadryl is an antihistamine that reduces the effects of natural chemical histamine which has the symptoms of sneezing, itching, watery eyes, and runny nose [30]. When performing the challenge, teenagers challenge each other to take large doses of Benadryl to trigger hallucinations. This will lead to fatal consequence in that taken in large enough doses of drug will health problem and even death, although right number of pills only cause

people sleepy. According to the US Food and Drug Administration, the health problem caused by too much medication includes serious heart problems, seizures, coma or even death [31].

B. Inappropriate video content

In addition to the text appear in the TikTok video and comment, the video content will also contain inappropriate things, which are even more difficult to be detected. For example, there was another trend in TikTok called the Silhouette challenge. In this trend, many female users will pose sexy in the video, and some will dance to the music rhythm. The mysterious shadow that seduces and cleverly shows off the body lines is shown with the backlit red light [31]. In some videos, underage users come across swearing, scantily clad adults, and suggestive dancing in TikTok. Many people even take advantage of this trend to edit the clip to hunt for nude images from TikTok's underage users.

Since these videos seem to have no association with nude trading, TikTok has failed to moderate them. For instance, 'Tradefortrade', a hashtag once spread in TikTok, connects to a profile with a male stroking themselves suggestively [32]. A comment on that post is, in fact, associated with trading nude images. One user's video asked for nude photos. Finally, the accounts are found to be connected to a community with lots of accounts centred around for nudity [32].

Additionally, there are some popular hashtags linked to dangerous behaviors like self-harm, suicide, and jumping off a building. Even choosing to filter out inappropriate content, one option in TikTok's Digital Wellbeing feature, cannot eliminate all the dangerous things. The filters can never catch everything since the hashtags change frequently with new creative spelling and hidden words or images can easily make the inappropriate content bypass filters. Additionally, hashtags can be edited to target a particular community, attributed to that the hashtag can help the videos send to people who have shown interest in it. When people swipe and watch the videos on the For You Page, the hashtag with specific words or content gradually reaches the particular community.

3.2 Challenges when solving algorithm bias

There are reigniting debates about what TikTok need to do when facing the ethical problem. It is critically important for TikTok to suppress content that involves violence, scamming, pornography, flatulence, and weighing in facts, high-quality content like news [20]. To reach this goal, TikTok is trying to create a border control frame needs to be defined beyond quantifiable model objectives (Content Audit system) [20]. However, challenges exist when solving algorithm bias. Rather than the quantifiable objectives in the recommendation system, such as click-through rate, reading time, likes, comments, and reposts, there are many things cannot be easily measured. This huge chunk of factors that cannot be easily evaluated or checked by the system cause many challenges when solving algorithm bias. This is due to that these intangible objectives, such as the content without hashtags or clear signals, cannot be identified by quantifiable indicators.

3.2.1 Content moderation

As mentioned before, one of the main challenges is the insufficient of content moderation, both text moderation and video moderation. Content moderation is the most used method to remove inappropriate content in recommendation system. However, content moderation is hard and cannot be perfect. TikTok must put a lot of work into making sure the algorithms catch any objectionable content. Content moderation includes pre-moderation and post-moderation. This means the content can be vetted before being posted or can be analyzed shortly after being posted.

Content moderation should not harm user interactions. When choosing to pre-review or post-review, TikTok needs to consider that having user videos and comments appear quickly is important to ensure a lively exchange of ideas and opinions, which is the social media platform's ultimate goal. If the likelihood of receiving problematic content is low, post-moderation may suffice for most of the

content [33]. Nevertheless, in TikTok, the likelihood of receiving problematic content is high.

Without completely pre-moderation, there are chances for malicious people to pose the worst content. Even with adequate post-moderation, either by using machine learning systems, or reviewing by human moderators, the worst content will get chances to be spread in ten seconds, ten minutes, or ten hours. Even within a short period, the power of spreadability cannot be alleviated. Cutting the video, editing it within seemingly harmless material, putting filters on it, or distorting the audio are all ways to spread the content [34]. Millions of variations for each unwanted word, including letter replacements, character insertions, phonetic variations, leetspeak, obfuscation, hyphenation, special characters, embeddings, etc, are hard to be eliminated. The possibility may be small, but once an inappropriate video gets spread, the impact is big. Machine learning and algorithm is now advancing but it can never be perfect, even in the future [35].

Therefore, post-moderation alone will not be timely and effective. The hybrid model is adopted to solve the problem to some extent. Ideally, this model includes software that blocks videos or comments from the moment they first appear online until reviewers evaluate them and make a final decision on whether to remove them. Yet even if the software is used to highlight questionable posts, TikTok should aim to read all videos and comments. There are not completely fixed and uniform standards to measure, and even personal feelings will be involved sometimes. Many people have said that the same video is deleted at first, but when posting it again, the video successfully pass verification.

3.2.2 No requirement for the video to go viral

As mentioned in the previous part, the recommendation system features the videos based on the users' interests. This means that a mass number of followers is not the requirement for the video to go viral. Although TikTok will ban the users who have spread inappropriate content in the past. There are ways that new users also spread these things. As TikTok has moved from local to global, more users are present, a larger possibility of having malicious people is present, and a bigger influence the content of videos will have.

The auto-play recommendation system in the For You Page making it easy to let thousands of users watch the video in just a few minutes. Without the limitation of based on who you follow, how many people you are following, TikTok finds it far more difficult to solve the ethical problem than other news feeds and social networking. For the users, after watching a video contain inappropriate material or disturbing content, the For You Page will start to recommend more to them again and again, even when they do not leave a single comment, interact with others, or follow the creator.

Furthermore, live streaming, a way to simultaneously record and broadcast in real-time to all the viewers around the world, is even harder to moderate. With an internet enabled device, such as a smartphone or tablet, and the broadcast platform, TikTok, almost any users can share anything in live streaming, which means the viewers will be exposed to any content the live streamer shows. The content in live streaming is even more difficult to moderate than videos and comments.

3.2.3 Migrate across to other platforms

As mentioned before, the suicide video once went viral in TikTok. The representative of TikTok once responded to the ethical problem:

"Our systems have been automatically detecting and flagging these clips for violating our policies against content that displays, praises, glorifies, or promotes suicide. We are banning accounts that repeatedly try to upload clips, and we appreciate our community members who've reported content and warned others against watching, engaging, or sharing such videos on any platform out of respect for the person and their family" [12].

With just a few taps users and reviews can remove the video showing disturbing content. Videos can be re-uploaded easily and migrated across other platforms quickly. To minimize the chance of detection

by the content moderation machine learning system, once the video is uploaded, bad actors are prone to download it, modulate it, and share it across multiple platforms like Reddit, Instagram, Facebook, and more [18]. It can very quickly become a cross-platform problem affecting many social media sites and thus millions of people around the world [35]. When the inappropriate videos are out in the wild and shown in other platforms, it will far more difficult for TikTok to remove them completely. Some people even move onto another app after making contact in TikTok and create bigger community to do illegal things. In this case, it will be hard to not only avoid the footage on For You Page when people scrolling through it, but also limit the spread of videos in only one platform.

4. Data governance and management solution

4.1 Solutions and effect

TikTok has announced a set of Community Guidelines to strengthen its existing policies focusing on the community's well-being and relieve social pressure. In the Community Guidelines, TikTok states how it processes inappropriate content and some other reactions to the ethical problem.

According to TikTok

We will remove any content – including video, audio, livestream, images, comments, and text – that violates our Community Guidelines. Individuals are notified of our decisions and can appeal if they believe no violation has occurred. We will suspend or ban accounts and/or devices that are involved in severe or repeated violations; we will consider information available on other platforms and offline in these decisions. When warranted, we will report the accounts to relevant legal authorities [36].

This Community Guidelines made it clearer about what kinds of content is banned in TikTok, but many users and reports were still criticizing TikTok, on the ground that this do not completely solve the problem. On February 24, 2021, TikTok published its latest transparency report. The report covers from the beginning of July to the end of December. According to this report, 89132938 videos were removed globally for violating the Community Guidelines. Moreover, TikTok states that it identified and removed 92.4% before a user reported them, 83.3% before they received any views, and 93.5% within 24 hours of being posted [37].

But removing videos like this will never completely ensure the health of the community. As stated in the report, there is no way to delete videos every single time. We can't deny TikTok's efforts and changes, but we can't completely solve the problems and challenges mentioned above. What TikTok needs is to make sure that they can't send the video out and spread it before it has an impact.

Another way to address the problem is to protect vulnerable groups, especially the underage users. TikTok has long had a series of protection and prohibition measures for underage users and children under the age of 13. For TikTok, it is also a way to deal with social criticism and skepticism. Children and adolescents are very vulnerable to the inappropriate content in TikTok. According to TikTok, registration for users who are below the age of 13 is not allowed.

In its Community Guidelines, TikTok claims "users must meet the minimum age requirements to use TikTok, as stipulated in our Terms of Service. When underage account holders are identified, TikTok will remove those accounts" [36]. It's true, however, that TikTok's primary audience is children (including those under 13) and teenagers. The fact is that it is easy for children and teenagers to register since the age verification process is so relaxed that only self-declaration is required [38]. Of the viral video-sharing app's 49 million daily users in the United States in July, about 18 million were 14 or younger, according to data obtained by The New York Times [39].

4.2 Grow in strict supervision

Under regulatory scrutiny, TikTok is facing threats of being banned in the US and many other countries as for data privacy and immoral content. TikTok was temporarily banned in Indonesia in 2018 and in Pakistan in 2020 [40]. In recent years, some countries or regions, especially the European Union, have also stepped up supervision and regulation of social me-

dia platforms in many aspects such as data protection and cyber hazards. Other countries have also imposed heavy fines and even prosecution on social media executives to effectively remove harmful content [41]. For example, in 2019, Sharing of Abhorrent Violent Material Act is passed in Australia. Technology executives can be sentenced to up to three years in prison and fined up to 10 percent of a company's global turnover [41]. The development of these legislation has put pressure on platforms to become more self-policing in terms of content censorship. Rather than ban TikTok, the regulatory plan is more efficient in that it offers a long-term, rules-based approach to addressing concerns about this social media platform.

5. Conclusion

5.1 TikTok data ethical dilemmas are inevitable

The problem of "immoral" content raises a kind of difficulty for the freedom of speech and is a dilemma that can hardly be solved [42]. The TikTok data ethical dilemma is that the Internet makes it easy for bad actors, ranging from trolls to spammers to malicious hackers, to deter or frustrate speech within online channels [43]. Nowadays, there are many problems and challenges in solving ethical problems. Even though both Tiktok and the government have taken steps to ensure the safety of users and the well-being of the community, there are still some potential problems. Many measures are being taken to mitigate the problem as much as possible, but they are fundamentally unavoidable. As long as there are people on Tiktok who want to take advantage of bad information or profit from it, there will be more ways for them to circumvent censorship and laws, and there will always be a certain chance that the recommendation system will spread the content, because enough hiding can make it past the screening and surveillance system. In general, the existence of these moral problems is undeniable and inevitable. The TikTok data ethical dilemma now becomes a permanent social problem sewn into the logic of the Internet [42]. Minimizing harm and potential problems is the best that Tiktok, the government, and users can do.

5.2 Three ways can help

Although the TikTok ethical dilemma is inevitable, there are three ways to help.

Firstly, for the users, protecting themselves is of great importance since it is the most direct and efficient way to maximize their security and privacy. Ensuring the TikTok is private or not allowing other users to find you and interact with you essentially means that others cannot follow, like, comment on, duet with you, find your account via a search engine, access to your videos and likes without approval. This can be done easily by turning the option on in the Safety section of the Privacy page, attributed to that TikTok share users' contents by featuring them on the For You Page by default. By blocking interactions and preventing others from viewing your content, the possibility that you receive inappropriate content or harassment under your videos and messages will be lower. Additionally, using restricted mode for children's accounts is another effective way. Restricted mode which stops inappropriate content cannot eliminate all the contents that are not suitable for underage users, such as false or misleading information, violence, sexually explicit material, hateful or offensive material, but it can help to some extent to alleviate the problem of seeing those dangerous, misleading content.

Secondly, government plays an indispensable role in managing social media. As mentioned before, banning social media is not the primary solution. For instance, since 2020, the UK government has outlined new powers for the media regulator Ofcom to regulate social media platforms [44]. Rather than rely largely on the self-governance of those platforms and let them make their own rules about removing inappropriate content, the government now forces them to ensure the harmful and objectionable content is removed in the first place. This means companies are responsible for this as well rather than only the person who posted it is at risk of prosecution [44].

Finally, it is a matter of fact that social media companies should be responsible for the inappropriate content and should aim at removing all the contents

that will potentially cause negative repercussions. In sharp contrast, the users, those bad actors in particular, won't get rid of their responsibility. The anonymity and free speech should not be the excuses for users to say whatever they want. Not only do social media companies need to be regulated, but the public should also be limited by laws. Legal restriction on content censorship standards is the prerequisite of eliminating the inappropriate contents but not removing them.

5.3 Further study aspects

The study suggests that future research should focus on how social software and short video apps deal with inappropriate information. Nowadays, many ethical issues are unavoidable, and there are still many potential problems and challenges. How to deal with and deal with these problems from the aspects of the software itself, government regulation, self-protection, and so on is very important. At the same time, it is hoped that there will be more profound research on how to improve the video recommendation system. Through such research, this software can better bring positive effects to users and eliminate inherent hidden dangers and dangers. At the same time, for a better network environment as well as the protection of teenagers' network use

The study suggests that there should be more research into the moral problems of TikTok in addition to its inappropriate content. It cannot be denied that TikTok has contributed in many ways and brought convenience and entertainment to people. But criticism of TikTok has been widespread across platforms and places. Only by further studying the remaining aspects of the problem, can we finally achieve the optimization of this social software as far as possible.

The study suggests more research on other social software in the future. Other social software also has some ethical issues to some extent. More research can be done on the problems that exist in other aspects as well as TikTok. Only by discovering these problems can we better improve them.

References:

1. Y. What Are Computer Algorithms, and How Do They Work? How-To Geek. (2016. September 28). URL: https://www.howtogeek.com/howto/44052/htg-explains-what-are-computer-algorithms-and-how-do-they-work
2. Mosaic A. K. The Powerful Equations That Explain The Patterns We See In Nature. Gizmodo Australia. (2020. August 25). URL: https://www.gizmodo.com.au/2014/08/the-powerful-equations-that-explain-the-patterns-we-see-in-nature
3. Sweigart A. How Does Compression Work? – The Invent with Python Blog. The Invent with Python Blog. (2012. August 17). URL: https://inventwithpython.com/blog/2012/08/17/how-does-compression-work
4. Note M. Managing Image Collections: A Practical Guide (Chandos Information Professional Series) (1st ed.). Chandos Publishing. (2011).
5. Tragakes E. Economics for the IB Diploma with CD-ROM (2nd ed.). Cambridge Univ. Press. (2011).
6. Cahn A., Alfeld S., Barford P., & Muthukrishnan S. An empirical study of web cookies. In: Proceedings of the 25th international conference on world wide web. 2016. – P. 891–901. International World Wide Web Conferences Steering Committee.
7. Ghose A., & Yang S. An empirical analysis of search engine advertising: Sponsored search in elec- tronic markets. Management Science, – 55(10). 2009. – P. 1605–1622.
8. DeAngelis S. E. F. S. Artificial Intelligence: How Algorithms Make Systems Smart. (2015, August 7). WIRED. URL:https://www.wired.com/insights/2014/09/artificial-intelligence-algorithms-2
9. Gibbs S. Gmail does scan all emails, new Google terms clarify. The Guardian. (2020, July 1). URL: https://www.theguardian.com/technology/2014/apr/15/gmail-scans-all-emails-new-google-terms-clarify

10. Kavenna J. Shoshana Zuboff: 'Surveillance capitalism is an assault on human autonomy.' The Guardian. (2019, October 29). URL: https://www.theguardian.com/books/2019/oct/04/shoshana-zuboff-surveillance-capitalism-assault-human-automomy-digital-privacy
11. Finley K. Wanna Build Your Own Google? Visit the App Store for Algorithms. Wired. (2014 b, August 11). URL: https://www.wired.com/2014/08/algorithmia
12. Congressional Research Service. TikTok: Technology Overview and Issues. (2020, January). URL: https://www.everycrsreport.com/files/2020–10–01_R46543_657b0e11b8b8cc8e9dfa307a961025825318883e.pdf
13. D. What is TikTok? Why Is It So Popular? Brandastic. (2020, November 27). URL: https://brandastic.com/blog/what-is-tiktok-and-why-is-it-so-popular
14. TikTok Revenue and Usage Statistics (2021). (2021, August 4). Business of Apps. URL: https://www.businessofapps.com/data/tik-tok-statistics/
15. Blacker A. Worldwide & US Download Leaders 2020. Apptopia. (2021, January 7). URL: https://blog.apptopia.com/worldwide-us-download-leaders-2020
16. Geyser W. TikTok Statistics – Revenue, Users & Engagement Stats (2021). Influencer Marketing Hub. (2021, July 8). URL: https://influencermarketinghub.com/tiktok-stats
17. Mohsin M. 10 TikTok Statistics You Need to Know in 2021 [March data]. Oberlo. (2021, July 1). URL: https://www.oberlo.com/blog/tiktok-statistics
18. Weimann G., & Masri N. Research Note: Spreading Hate on TikTok. Studies in Conflict & Terrorism. Published. (2020). URL: https://doi.org/10.1080/1057610X.2020.1780027
19. Zhao Z. Analysis on the "Douyin (Tiktok) Mania" Phenomenon Based on Recommendation Algorithms. (2020). E3S Web of Conferences 235, 03029 (2021). NETID2020. Published. URL: https://doi.org/10.1051/e3sconf/202123503029NETID2020
20. Wang C. Why TikTok made its user so obsessive? The AI Algorithm that got you hooked. Medium. (2020, June 7). URL: https://towardsdatascience.com/why-tiktok-made-its-user-so-obsessive-the-ai-algorithm-that-got-you-hooked-7895bb1ab423
21. T. How TikTok recommends videos (2020b, November 5).#ForYou. Newsroom | TikTok. https://newsroom.tiktok.com/en-us/how-tiktok-recommends-videos-for-you
22. Matsakis L. How TikTok's "For You" Algorithm Works. Wired. (2020, June 18). URL: https://www.wired.com/story/tiktok-finally-explains-for-you-algorithm-works
23. Han B., & Butler E. In the Swarm: Digital Prospects (Untimely Meditations). The MIT Press. (2017).
24. Warnock C. TikTok Works to Take Down Disturbing Suicide Video as Users Share Warnings. Heavy. Com. (2020, September 8). URL: https://heavy.com/news/2020/09/tiktok-viral-suicide-video
25. Associated Press. TikTok says coordinated attack behind suicide clip uploads. Washington Post. (2020, September 22). URL: https://www.washingtonpost.com/business/technology/tiktok-says-coordinated-attack-behind-suicide-clip-uploads/2020/09/22/5c0b9a98-fcdb-11ea-b0e4–350e4e60cc91_story.html
26. Hosn J. TikTok Video Comments – David Dobrik's top videos: Scraped video comments on David Dobrik's top videos on TikTok. Kaggle. (2020).
27. Databricks. Apache Spark™ – What is Spark. (2020, April 14). URL:https://databricks.com/spark/about
28. Leander K. M. Critical literacy for a posthuman world: When people read, and become, with machines. British Educational Research Association. (2020, July 1). URL: https://bera-journals.onlinelibrary.wiley.com/doi/abs/10.1111/bjet.12924

29. Statista. TikTok MAU user ratio in the U.S. 2019, by age group & gender. (2021, June 7). URL: https://www.statista.com/statistics/1095196/tiktok-us-age-gender-reach/#statisticContainer
30. Benadryl. Drugs.Com. (2020, December 8). URL: https://www.drugs.com/benadryl.html
31. Nguyen B. Don't let your kids try these 9 dangerous TikTok trends. CyberPurify. (2021, August 19). URL: https://cyberpurify.com/knowledge/9-dangerous-tiktok-trends
32. Cox J. TikTok, the App Super Popular With Kids, Has a Nudes Problem. Vice. (2018, December 6). URL: https://www.vice.com/en/article/j5zbmx/tiktok-the-app-super-popular-with-kids-has-a-nudes-problem
33. Pre- and post-moderation. (2019, June 18). Ontheline. URL: https://newsrooms-ontheline.ipi.media/measures/pre-and-post-moderation/
34. Productions B., Productions B., Productions B., Productions B., Productions B., Productions B., & Productions B. Blog: ANALYSIS – TikTok suicide video: it's time platforms collaborated to limit disturbing content – AdNews. (2020, September 9). Best Soundcloud Rappers 2019. URL: https://btoktiktok.com/2020/09/09/blog-analysis-tiktok-suicide-video-its-time-platforms-collaborated-to-limit-disturbing-content-adnews
35. ABC News. TikTok suicide video: It's time social media platforms collaborated to limit disturbing content. (2020, September 9). URL: https://www.abc.net.au/news/2020–09–09/why-so-hard-tiktok-remove-disturbing-content-suicide-video/12643832
36. TikTok Community Guidelines. (n.d.). TikTok. Retrieved August 21, 2021. From URL: https://www.tiktok.com/community-guidelines?lang=en
37. A. (2021). TikTok Transparency Report 2020 H1. TikTok. https://www.tiktok.com/safety/resources/transparency-report-2020–1?lang=en
38. BEUC The European Consumer Organisation. (2021). TikTok without filters. BEUC The European Consumer Organisation. Published. URL: https://www.beuc.eu/publications/beuc-x-2021–012_tiktok_without_filters.pdf
39. Zhong R., & Frenkel S. A Third of TikTok's U. S. Users May Be 14 or Under, Raising Safety Questions. The New York Times. (2020, September 18).URL: https://www.nytimes.com/2020/08/14/technology/tiktok-underage-users-ftc.html
40. Wang D. J. From Banning to Regulating TikTok: Addressing concerns of national security, privacy, and online harms. The Foundation for Law Justice and Society. (2021). Published. URL: https://www.fljs.org/sites/default/files/migrated/publications/From%20Banning%20to%20Regulating%20TikTok.pdf
41. Criminal Code Amendment (Sharing of Abhorrent Violent Material) Act 2019. (2019). Federal Register of Legislation. URL: https://www.legislation.gov.au/Details/C2019A00038; URL: https://www.kaggle.com/jhosn13/tiktok-video-comments-david-dobriks-top-videos
42. Langvardt K. Regulating Online Content Moderation. The Georgetown Law Journal. (2018). Published. URL: https://www.law.georgetown.edu/georgetown-law-journal/wp-content/uploads/sites/26/2018/07/Regulating-Online-Content-Moderation.pdf
43. Lidsky L. B. Public Forum 2.0. UF Law Scholarship Repository. (2011). Published. URL: http://scholarship.law.ufl.edu/facultypub/155
44. Team B. R. C. Social media: How do other governments regulate it? BBC News. (2020, February 12). URL: https://www.bbc.com/news/technology-47135058

Contents