**Vineeth Karumanchi**

2020

Leveraging Product Characteristics for Online Collusive Detection in Big Data Transactions

**List of Abbreviations**

IP - Internet Protocol

IT - Information Technology

AI - Artificial Intelligence

ML - Machine Learning

DMLC - Distributed Machine Learning Community

CLI - Command Line Interface

TP - True Positive

FP - False Positive

TN - True Negative

FN - False Negative

CVE - Common Vulnerabilities and Exposures

IQ - Intelligent Quotient

IBM - International Business Machines Corporation

HTML - Hypertext Markup Language

CSV - Comma Separated Values

iOS- iphone Operating System

# Abstract

Online fraudulent transactions are a big concern for the e-business platform. As big data technology evolves, e-commerce users are always evaluated by sellers according to the reputation scores offered by the platform. The reason why sellers like high reputation scores is that a high rating always gives sellers high profits. Through partnerships, fraudsters can achieve higher profile scores and this will attract more potential customers. The e-commerce website plays a key role in detecting fake reputation information. E-commerce platforms continue to grow by adopting data mining practices. With the high development of the Internet of Things (IoT), big data plays a key role in financial society. Big data achieves economic growth across different domains. It supports the management and decision making capacity of the e-business by analyzing performance data. In online commerce, big data technology helps to give consumers a reasonably healthy reputation system that enhances the shopping experience. This paper puts forward a theoretical framework for extracting the characteristics of fraudulent transactions, including personal and transaction indicators. It has two product characteristics: product type and product behavior. Both features increase the accuracy of fraud detection. The real-world dataset is used to test the effectiveness of indicators in the detection model, which comes forward to detect fraudulent transactions from law firms.

**Keywords:** Big data, collusive detection, IoT, Leveraging, Product Characteristics, economic growth, bagging algorithms, boosting algorithms.

# TABLE OF CONTENTS

[Abstract 3](#_Toc60869553)

[TABLE OF CONTENTS 4](#_Toc60869554)

[Chapter 1: Introduction 8](#_Toc60869555)

[1.1 Introduction 8](#_Toc60869556)

[1.2 Background 8](#_Toc60869557)

[1.3 Research Rationale 10](#_Toc60869558)

[1.4 Research Aim 11](#_Toc60869559)

[1.5 Research Questions 12](#_Toc60869560)

[1.6 Significance 12](#_Toc60869561)

[1.7 Conclusion 13](#_Toc60869562)

[Chapter 2: Literature Review 15](#_Toc60869563)

[2.1 Introduction 15](#_Toc60869564)

[2.2 Empirical Study 16](#_Toc60869565)

[2.3 Tools and Models 18](#_Toc60869566)

[2.4 Literature Gap 19](#_Toc60869567)

[2.5 Conceptual Framework 20](#_Toc60869568)

[2.6 Conclusion 22](#_Toc60869569)

[Chapter 3: Methodology 23](#_Toc60869570)

[3.1. Proposed Evaluation Metrics **Error! Bookmark not defined.**](#_Toc60869571)

[3.1.1. Accuracy **Error! Bookmark not defined.**](#_Toc60869572)

[3.2. Data Pre-processing **Error! Bookmark not defined.**](#_Toc60869573)

[3.3. Evaluation of Metrics **Error! Bookmark not defined.**](#_Toc60869574)

[3.4. Data Exploration and Data Analysis **Error! Bookmark not defined.**](#_Toc60869575)

[Chapter 4: Results and Analysis 25](#_Toc60869576)

[4.1 Data Exploration 25](#_Toc60869577)

[4.2 Data Analysis 30](#_Toc60869578)

[4.3. Feature Selection **Error! Bookmark not defined.**](#_Toc60869579)

[4.4. Machine Learning 32](#_Toc60869580)

[Chapter 5: Conclusions and Future work 51](#_Toc60869581)

[5.1. Conclusion 51](#_Toc60869582)

[5.2. Research Limitations **Error! Bookmark not defined.**](#_Toc60869583)

[5.3. Future Work 52](#_Toc60869584)

[References 53](#_Toc60869585)

**TABLE OF FIGURES**

[Figure 1: Data pre-processing Diagram **Error! Bookmark not defined.**](#_Toc60910997)

[Figure 2: Class count plot **Error! Bookmark not defined.**](#_Toc60910998)

[Figurec 3: Class count plot after Class Balancing **Error! Bookmark not defined.**](#_Toc60910999)

[Figure 4: Data Feature analysis **Error! Bookmark not defined.**](#_Toc60911000)

[Figure: Supervised Machine learning **Error! Bookmark not defined.**](#_Toc60911001)

[Figure: Unsupervised machine learning **Error! Bookmark not defined.**](#_Toc60911002)

[Figure: Reinforced Machine Learning **Error! Bookmark not defined.**](#_Toc60911003)

[Figure: Evaluation metrics for CATBoost Classifier **Error! Bookmark not defined.**](#_Toc60911004)

[Figure: Precision vs. Recall plot for CATBoost Classifier **Error! Bookmark not defined.**](#_Toc60911005)

[Figure: Accuracy Matrix for CATBoost Classifier **Error! Bookmark not defined.**](#_Toc60911006)

[Figure: True positive rate vs. false positive rate for CATBoost Classifier **Error! Bookmark not defined.**](#_Toc60911007)

[Figure: Evaluation metrics for Gradient Boost Classifier **Error! Bookmark not defined.**](#_Toc60911008)

[Figure: Precision vs. Recall plot for Gradient Boost Classifier **Error! Bookmark not defined.**](#_Toc60911009)

[Figure: Accuracy Matrix for Gradient Boost Classifier **Error! Bookmark not defined.**](#_Toc60911010)

[Figure: True positive rate vs. false positive rate for Gradient Boost Classifier **Error! Bookmark not defined.**](#_Toc60911011)

[Figure: Evaluation metrics for XGBoost Classifier **Error! Bookmark not defined.**](#_Toc60911012)

[Figure: Precision vs. Recall plot for XGBoost Classifier **Error! Bookmark not defined.**](#_Toc60911013)

[Figure: Accuracy Matrix for XGBoost Classifier **Error! Bookmark not defined.**](#_Toc60911014)

[Figure: True positive rate vs. false positive rate for XGBoost Classifier **Error! Bookmark not defined.**](#_Toc60911015)

[Figure: Evaluation metrics for ADABoost Classifier **Error! Bookmark not defined.**](#_Toc60911016)

[Figure: Precision vs. Recall plot for ADABoost Classifier **Error! Bookmark not defined.**](#_Toc60911017)

[Figure: Accuracy Matrix for ADABoost Classifier **Error! Bookmark not defined.**](#_Toc60911018)

[Figure: True positive rate vs. false positive rate for ADABoost Classifier **Error! Bookmark not defined.**](#_Toc60911019)

[Figure: Comparative Analysis for the Boosting Algorithms **Error! Bookmark not defined.**](#_Toc60911020)

[Figure: Evaluation metrics for Extra Trees Classifier **Error! Bookmark not defined.**](#_Toc60911021)

[Figure: Precision vs. Recall plot for Extra Trees Classifier **Error! Bookmark not defined.**](#_Toc60911022)

[Figure: Accuracy Matrix for Extra Trees Classifier **Error! Bookmark not defined.**](#_Toc60911023)

[Figure: True positive rate vs. false positive rate for Extra Trees Classifier **Error! Bookmark not defined.**](#_Toc60911024)

[Figure: Evaluation metrics for Random Forest Trees Classifier **Error! Bookmark not defined.**](#_Toc60911025)

[Figure: Precision vs. Recall plot for Random Forest Trees Classifier **Error! Bookmark not defined.**](#_Toc60911026)

[Figure: Accuracy Matrix for Random Forest Trees Classifier **Error! Bookmark not defined.**](#_Toc60911027)

[Figure: True positive rate vs. false positive rate for Random Forest Trees Classifier **Error! Bookmark not defined.**](#_Toc60911028)

[Figure: Evaluation metrics for Decision Trees Classifier **Error! Bookmark not defined.**](#_Toc60911029)

[Figure: Precision vs. Recall plot for Decision Trees Classifier **Error! Bookmark not defined.**](#_Toc60911030)

[Figure: Accuracy Matrix for Decision Trees Classifier **Error! Bookmark not defined.**](#_Toc60911031)

[Figure: True positive rate vs. false positive rate for Decision Trees Classifier **Error! Bookmark not defined.**](#_Toc60911032)

[Figure: Comparative Analysis for Boosting and Bagging algorithms **Error! Bookmark not defined.**](#_Toc60911033)

[Figure: Comparative bar plots for all the model evaluation **Error! Bookmark not defined.**](#_Toc60911034)

# Chapter 1: Introduction

# Introduction

Online fraudulent transactions are a big concern for the e-business platform. As big data technology evolves, e-commerce users are always evaluated by sellers according to the reputation scores offered by the platform. The reason why sellers like high reputation scores is that a high rating always gives sellers high profits. Through partnerships, fraudsters can achieve higher profile scores and this will attract more potential customers. The e-commerce website plays a key role in detecting fake reputation information. E-commerce platforms continue to grow by adopting data mining practices. With the high development of the Internet of Things (IoT), big data plays a key role in financial society. Big data achieves economic growth across different domains. It supports the management and decision making capacity of the e-business by analyzing performance data. In online commerce, big data technology helps to provide consumers with a reasonably healthy reputation system that enhances the shopping experience. This paper puts forward a theoretical framework for extracting the characteristics of fraudulent transactions, including personal and transaction indicators. It has two product characteristics: product type and product behavior. Both features increase the accuracy of fraud detection. The real-world dataset is used to test the effectiveness of indicators in the detection model, which comes forward to detect fraudulent transactions from law firms.

# Background

This is a big concern for the e-business platform in online fraudulent transactions. E-commerce users will always be able to predict the reputation scores offered by a platform developed by big data technology. A high rating always gives sellers high profits because the seller wants to get the high reputation that the platform offers. By gaining a high reputation fraud can become a community that attracts more potential customers. The e-commerce website plays a key role in detecting fake reputation information. It continues with the problems that e-commerce platforms are trying to solve by adopting data mining methods. Big data plays a key role in a financial society with a well-developed Internet of Things. Economic growth takes place across different domains using big data. Supports e-business distribution management and decision making skills by analyzing performance data. The main objective of this paper is to incorporate personal and transactional indicators by providing an affordable and healthy reputation system for online commerce, a conceptual framework that enhances the shopping experience and features with great data technology. Fraudulent transaction has two product features Product type and product character. The accuracy of fraud detection is determined by these two characteristics. The real-world dataset is used to test the effectiveness of indicators in the detection model, which proceeds to detect fraudulent transactions from legitimate ones.

Using the Internet of Things creates a large amount of data. Online commerce adopts relevant data such as data mining and machine learning to obtain valuable business information using big data technology. It boosts sales through the use of IoT, big data management and a healthy shopping platform environment. The popularity of online shopping has grown with the advent of big data technology, high efficiency and low internet cost. China Internet Network Information Center (CNIC) has 772 million online users in China, founded in 2017 by China Internet Development. Statistics and report published in 2018. Recently, the Tobao platform has more than one million fraudulent sellers each year, with over 500 million transactions and $ 10 billion.

Many online business websites solve this problem by providing a recommendation system or credit information system to help potential buyers distinguish a legitimate shopper. Most online shopping platforms like JD, DingDong, Yelp and Tobao use the reputation system to prevent fraudulent transactions. The e-business environment requires an impression system. Due to the geographical distance the buyer has to deliver the goods after paying online. This increases the risk. There are a few reputation systems in the historical transaction that show customer reviews. These systems play a key role in online transactions. Upon completion of their transaction, some reputation systems must give rival scores to rivals on both sides of the trade. Current rating systems may be the sum or average of the ratings collected. The system maintains the final score with negative scores minus the positive scores. In this way, Tobao uses the sum of the platform ratings. In general, potential buyers are more likely to shop with more reputable sellers. Significant scams are taking place in the online marketplace. Relevant mass data can be collected from large data generated by the IoT platform and valuable information can be obtained to improve service and profitability. A high rating always gives sellers high profits, so sellers prefer to get a high reputation. Substantial fraudulent activities have led to the temptations of financial gain and the difficulty of internet monitoring. Impression systems are always attacked by an illegal organization and do not reflect the reputation of the merchants. Finding infamous reputation scams for online shopping platforms is an important task because online buyers rely on the reputation system to assess sellers.

# Research Rationale

Online identification is hugely needed to identify illegal users based on the large amount of information and big data that IoT has acquired. However, the relationship between customers weakens the credibility of reputation systems but receives very limited attention. Systematic anti-fraud solutions are by no means rare, so limited resources are included in the anti-fraud field. (Tobao, 2019) is taking a keen interest in resolving fraudulent transactions. There have been some studies on the types of scams, the motives behind scams, the impact of fake websites and auction scams such as tea. Various methods have been put forward, including graph mining methods, decision perspectives, regression models, model testing and statistics. A detailed overview of methods, classical classification methods of neural networks, clustering, neighborhood diversity and methods of detecting scams has been discussed.

Our study papers cover three topics: (1) We introduce two new features of fraudulent transactions and link them with other user roles for fraud detection. Several elements of online transaction behavior were collected, including product type and product behavior. This is useful for other types of online combinations when developing Discovery models. Ways to get indicators including money laundering, tax evasion, smuggling and drug trafficking.

The reputation of mass sellers will increase significantly. A collective transaction that misleads buyers can be considered fraudulent. As a basic understanding of neo-classical economics, Homo economics sees people as self-serving and hostile. Sociologists have not confirmed the hypothesis of pure economic man; they wanted to reconcile Homo economics with the hypothesis of Homo sociology. However, the neoclassical economist Becker [7] suggested that criminals make rational decisions. Many scholars have argued that the homo-economic paradigm is crucial to survival success in certain aspects of life. Therefore, when the shortcomings of their crime are overcome, the perpetrators, as individuals who deliberately commit crimes, commit frauds.

Existing system: Large society plays a key role in economic society, as well as the high development of the Internet. Big data increases economic growth across different domains. Supports e-business distribution management and decision making skills by analyzing performance data. Online commerce enhances the shopping experience with great data technology that gives consumers an affordable and healthy reputation system. It is difficult to pinpoint the exact identity of an e-commerce partner because the features of the online shopping environment are virtual in nature. Achieving the product quality that buyers want is not always easy due to disproportionate information.

Suggested method: In our study, fraudulent transactions, aimed at enhancing reputation, refer to illegal transactions undertaken by illegal associations for profit. The term puppet buyer refers to the ID registered by the collective group, who has to make fraudulent deals with sellers who pay for increased reputation. Collective seller refers to a fraudulent attempt to obtain a high credit score illegally. The typical way to increase popularity is shown in Figure 1. Mass sellers: After paying for the service, the unauthorized company completes multiple joint transactions with the dummy distribution of the product to the buyer of many puppets. As a result, the positive ratings and comments rated by buyers of this toy is unpredictable.

# Research Aim

Significant scams are taking place in the online marketplace. Collect platform-related mass data and obtain valuable information to improve service and profitability from large data generated by IoT. The reasons for that High reputation always gives sellers high profits, sellers want to get high reputation. Through partnerships, fraudsters can achieve higher profile scores and this will attract more potential customers. It gives more profit to the illegal’s Vendors. Therefore, it forces them to make a collective bargain. The lure of financial gain and the difficulty of monitoring the Internet have led to significant fraud [4].

The reputation of the merchant is always being attacked by an illegal entity. The number will also increase as the e-business expands. Fraudulent transactions are on the rise. Online buyers will find out because they rely on the reputation system to evaluate sellers Inflation reputation fraud is an important liability for online shopping platforms. Online detection requires massive identification to find illegal users based on valuable information obtained through IoT and big data. However, reputation undermines the credibility of systems, but only very limited attention is paid to interaction between customers.

# Research Questions

The following research questions can be utilized for the collusive detection for the online transaction in the big data platform. The main objective of the study is to not use term fraud instead use Collusive which means involving secret or unlawful cooperation aimed at deceiving or gaining an advantage over others, which is what fraudsters do when they gain access to someone credit card.

1. What are the significant difference between benign participator and fraudster behavior?
2. What difference results were observed in boosting algorithm and bagging algorithm?

# Significance

One of our goals is to present a framework for obtaining counterfeit transaction letters on the e-business platform. There is no very effective way to identify fraudsters. There have been several attempts in this area. However, all available information is not collected from user data, i.e. fraud detection (Vizia et al., 2011) and mass fraud detection in online reviews (Chang, 2016).

We identify the costs and benefits of the collective organization and then select a set of features with two valid product components that are valid in distinguishing between beneficial and undesirable behavior. We use more information available by putting forward variables such as review length, product type and product behavior. In our study, there was a significant difference in behavior between the harmless partner and the cheater. The process of obtaining these indicators can be generalized to minor text classifications and other scams.

The online shopping platform appears much later than a foreign country in China. A limited resource is included in the fraud prevention area, so systematic anti-fraud solutions are by no means rare. Tobao is very interested in resolving fraudulent transactions. There have been some studies on the types of fraud [2], the motives behind fraud [6] and auction fraud.

Impact of counterfeit websites [25], Chai et al. [14] Studies the consequences of fraud detection in the online trading community, explicit behavioral evidence [3] and monitored studies to identify online auction collisions [6]. There is some research on fraud detection. Various methods have been proposed, such as graph mining methods [5], decision perspectives [2], [9], statistical methods [6], classical classification methods of neural networks, model testing [7] and clustering [15].

There are very different behaviors between a harmless partner and a fraudster. In mass transactions, newly registered counterfeit buyers actively participate in counterfeit transactions by shopping for virtual and cheap products. They are highly appreciated for their high rating status and detailed writing reviews. Further research may focus on developing the above model more universally. Various types of informational data from other e-business platforms can be used to test the effectiveness and cost of fraudulent transactions to test the effectiveness of the partnership identity model. What’s more, mass organizations are trying to improve their skills to avoid finding Tobao. Therefore, one direction of our next research is to develop a more favorable identification model.

# Conclusion

A new approach has been introduced in this paper to find out the collective behavior of consumers. We have identified new features of online transaction. These features are extracted through data processing and operate according to our detection model. This model can be used to distinguish between collective transactions and good reliability. There are three aspects to the contribution of our study:

1. We will introduce two new features of fraudulent transactions and link them with other user characters for fraudulent identification. Several aspects of online transaction behavior, including product type and product behavior are summarized. It is useful for other types of online combinations in developing detection models. The method of obtaining indicators can be generalized to identify other behaviors, including money laundering, tax evasion, smuggling, and drug trafficking.
2. We use the real-world dataset to test the feasibility of our innovation model.
3. We provide some suggestions on the platform of online e-commerce to protect the online reputation environment.

Introduction and Literary Survey Results:

(1) We introduce two new features of fraudulent transactions and link them with other user roles for fraud detection. Our method of obtaining indicators can be generalized to other collective behavior identities.

(2) We use the real-world dataset to test the feasibility of our innovation model.

(3) We provide some suggestions on the platform of online e-commerce to protect the online reputation environment. We also provide our relevant explanations. The process of obtaining these indicators can be generalized with short text classification and other types of fraudulent activities. There are very different behaviors between a harmless partner and a cheater. In mass transactions, newly registered counterfeit buyers actively participate in counterfeit transactions by shopping for virtual and cheap products. They are highly appreciated for their high rating status and description

Writing reviews. Further research may focus on developing the above model more universally. Various types of informational data from other e-business platforms can be used to test the effectiveness and cost of fraudulent transactions to test the effectiveness of the partnership identity model. What’s more, mass organizations are trying to improve their skills to avoid finding Tobao. Therefore, one direction of our next research is to develop a more favorable identification model.

# Chapter 2: Literature Review

# 2.1 Introduction

Fraud prevention and fraud detection are two types of approaches to prevent online scams. Focus on preventing fraud in the face of fraudulent behaviors, including personal identification systems and online credit transaction security systems. At the same time, fraudulent results [5] are used to identify fraudulent transactions. [31] A research summarizes research on fraudulent behavior. Puter Lloyd Al. [4] although many methods can be used to prevent fraudulent activities, e-commerce has been found to use a systematic approach. Maransatot al. [5] Logistics regression has been used to identify and diagnose the symptoms of credit fraud. [5] [7] the logistic regression model is used to detect online transactions on the Dingdong platform. For fraudulent research, the Rana & Baria Review [3] is a brief research on methods of detecting past frauds. In the past, research has focused primarily on fraud in accounting, auditing and financial activities. The increase in online transactions has led to an increase in fraudulent activity-relationships and the lack of online monitoring has made it easier to deceive sellers [6]. For a detailed overview of fraud detection data mining practices, please see [19]. Con artists should try to protect themselves from being identified by mimicking legitimate behaviors, making it difficult to measure collective behavior. Crowd-sourcing and human computing has been suggested to enhance current recognition capabilities. Before making contact with a crowdsourcing agent, you must first identify potential users who may be scammers. We will try to address this in our study. There is no better way to detect corporate scams and, as in previous methods [6], some valuable features have not been collected in mass fraud detection online reviews [15]. The price and popularity of the product and some control variables have been improved by the use of two new letters to identify open fraudsters. We use both product features to increase the accuracy of online mass innovation and use the real-world dataset to test the feasibility of our innovation model. Based on the practical result, we give some hints to protect the online reputation-environment of the platform approach in online e-commerce which includes short text classification technology to gather more valid information. The combination has developed an identity model. We strive to make full use of available information, including review length, product type and product form. Before taking effective measures to prevent fraudulent transactions, potential customers of fraud must be identified in advance. This is what we are trying to address in our study.

# 2.2 Empirical Study

In addition to customer information provided by businesses, businesses may collect information about customer online behavior using cookies and click-stream analytics that do not require the conscious participation of customers (Rust, Cannon, Peng 2018). Data is less technology

Collection, storage and retrieval is lower than ever before. This has led to efficient and cost-effective data mining methods and data warehousing technology, allowing marketers to better analyze and target customers (Markoff 2019, Richards 2018). Facilitates a network environment that collects and codes customer information

Effectively distribute or sell the information collected, thereby integrating the complete and integrated profile of customers and parts of customer information that may look different to improve their behavior (Romble 2017). Therefore, consumers will not be completely upset if they start looking at online marketing efforts with the Reserve, especially as marketing methods on the Internet give consumers less control over their information (Franz, Pitta, Fritz, 2017). Consumers hope that retailers will provide a unique method of personalized purchase information to facilitate transactions. In some cases, sharing personal information with businesses and other organizations they trust has proven acceptable, and some users see all requests and uses of personal information as an invasion of privacy (Domer & Gross 2013). In addition, many consumers are willing to give up some privacy to participate in the consumer community. However, consumers are more likely to believe that retailers will benefit from their costs if they sell information about personal purchasing practices to other sellers (Graf, Hormone, 2016). It is important to recognize that this marketing exchange has two sides, and that both sides have many rights and responsibilities; And the moral issues of human dignity, value, autonomy, province, anonymity, security and protection (Mascarenhas, Keshavan, Bernachi, 2017) need to be considered.

After the transaction, the online trading platform provides an evaluation system to gather the reputation of the participants. Participants rate their competitors after each transaction [7]. As online reputation scores have become more influential in making purchasing decisions for online shoppers, cybercriminals have a strong financial incentive to participate in mass bargaining [3]. Mass fraud has emerged as a major threat to the e-commerce platform. The exaggerated design of the online reputation system creates a sanctuary for mass fraudsters. Con artists use the registration facility to deceive buyers. Therefore, developing an effective method to maintain a healthy reputation system is a very crucial task. Currently, random transactions on the online reputation systems of e-commerce platforms are systematically carried out by illegal organizations. It is very difficult to find such collective behaviors because they mimic the routine of bad transactions which highlights the typical process of fraudulent transactions on e-business platforms. Collective organization is an organized association of criminals. First, a fraudulent shopkeeper points out a named inflation charge to the coalition agency and pays the charge. The company begins to hire toy buyers and hand over responsibilities.

In the context of new technologies, there have always been legal and social crises created for marketers. Consumers generally do not believe that marketers are interested in consumer privacy issues, and have negative perceptions of marketers trying to record more personal information (Graf & Hormone, 2017). At the same time, consumers agree that even though some marketers already know a lot, consumers know more about themselves and their catalogs and promotional messages will be more useful (Phelps et al., 2018). The lack of a clear framework to manage this data can lead to the misuse of such technology and the creation of such crises. Such abuse can have a large strategic impact on an organization, destroying its reputation and limiting the credibility it can develop with respect to customers, employees, channel members and competitors (Bloom, Milne, Adler, 2009). Lack of knowledge raises consumer concerns and is essential in preventing such crises. Information decision-making plays a key role in the fair dissemination of knowledge (Kulnan, Armstrong, 2017).

Bloom, Milne and Adler (2014) identified four problems that marketers face when using new technology:

1. Engage in mass communication

2. Maintain illegal necessary facility

3. Store or transfer inaccurate or malicious information

4. Violate privacy rights of individuals. Although these four dimensions have been identified as serious issues that marketers need to be aware of when adopting new information technology, they can be further strengthened if we apply the principles in an online environment, because in order to initiate online transactions, users have created large databases containing their population and purchasing habits. Today's technology allows for more advanced data collection and invasion of privacy (Graf & Hormone, 2017). The growing use of such data collection technology is one of the major challenges facing data marketers today. Unlike consumer surveys (telephone, mall-intercept and mail), which can only measure the declared purchasing intentions and attitudes of consumers, database marketing and automated data capture systems allow sellers to record real buying behavior (ibid). It has been argued that databases are created through voluntary contributions from customers who support data collection, although this view ignores the fact that providing personal data is the only way to proceed with a transaction. Similarly, the process of obtaining data is called ‘semi-voluntary’.

# 2.3 Tools and Models

The fraudulent identification model uses logical regression because the logistics regression model has the advantage over the number of features and the model is enhanced with limited or non-existent features. Since thirteen percent is used to detect financial fraud, the logistics regression model is placed on the edge of the data mining tool [1]. Although not a reflection of linear variables, this model is widely used in mass estimation [4]. The logistics model and the logistics function are widely used to create a statistical model binary dependent variable in its basic form which comprises of Binary regression, in the form of logistic regression.

Calculates the parameters of the logistics model; Regression analysis is used. Mathematically, the binary logistics model consists of a dependent variable with two possible values; these are represented by an index variable, where both the values ​​0 and 1 are labeled. The label used by the collective bargaining logistics model. In the practical study, we use a real-world dataset provided by the Tobao platform. Our data was obtained in accordance with the disclosure agreement. Duplicate accounts contain data related to fraudulent transactions and related transaction records. We used a self-developed web crawler program to capture bad commercial records. It contributes to the structure of our transaction network. We collect data about well-behaved customers associated with fraudulent sellers labeled using the platform. Finally, we get a lot of information about legitimate users who are identified as inappropriate users according to the platform. There were 170899 transactions, of which 2917 were fraudulent transactions (1.71%), 803 were corporate accounts and 23,401 were non-collective accounts which we used for analysis. After the data is cleaned, prepared, and the detection model is built, we use the Sci-kit lab package to achieve logical regression or the Python environment. Launched in 2007, Sci Kit-Lane has become a major machine learning library for Python. Supports four machine learning algorithms, including classification, regression, and reduction and clustering can be utilized for the research. It consists of three modules: feature extraction, data processing and model evaluation. Sklearn is an extension of Scipy based on the NumPy and matplotlib libraries. By using these modules, the efficiency of machine learning can be greatly improved. Scilearn has a great API that is popular at the academy. Scilearn already has several machine learning algorithms, including logical regression. At the same time, Schlern has a lot of built-in data sets that save time in owning and sorting data sets. We examine specific indicators by following the approach of logic regression. Significant slope emphasizes the need for these transitions because some variables, including the buyer's K-score, buyer's age, number of monthly transactions, review length, product price and seller reputation, are replaced with corresponding logarithmic equivalents.

# 2.4 Literature Gap

The literature review section has shown that several research works have been conducted in the field of fake review detection using machine learning techniques. Earlier measures were done by manual identification and simple visualization tools. The work explicitly focuses on the comparative analysis of all the regression algorithms of machine learning namelyLabelBagging algorithms, boosting algorithms and CATBoost algorithm. The literature survey discussed the use of support vector machine algorithms, use of neural networks, reinforcement learning and the time theory of the machine learning techniques.

# 2.5 Conceptual Framework

As a fraudster, you should consider two types of expenses. One is financial costs, the other one is behavioral costs. Behavioral costs refer to the amount the seller has to pay to complete the booked order. These effects include device effects and target effects suggested by Verhalen [6]. Target brings effects Expected successes, when performing instrumental activities with the goal of achieving the goal. Helps to create registration

Puppet buyer is involved in online business environment. All the behaviors done to falsely implement these false processes are done to extract monetary profits. Trading can be considered a target effect, including product selection, payments, rating levels and comments or as an advice. In general, it is very easy for a buyer to register a new user account. As far as we know, only a valid telephone number is required to obtain user identification. Users must enter the verification code sent to their registered phone numbers when setting up the account. Therefore, the cost is very low. However, it requires reasonable time and effort.S. Luo, s to develop or employ people to register computer programs. VAN: Product behavior features maintain multiple phone accounts for the online mass discovery of big data transactions. The best way to save money Place as many orders as possible with multiple sellers. There is a feature of fraudsters; the activity is also mentioned Bloom et al. Therefore, the first theory is presented below:

Hypothesis 1 (fraudulent consumer livelihood hypothesis): Fraudsters are involved in a lot of transactions those ones. To protect the interests of the buyer, most platforms use third party payment software to hold cash before confirming receipt of goods. This creates a financial expense for the collective entity caused by the time value of the money. When the buyer is satisfied with the product, they can click the "Confirm" button on the e-commerce website to complete.

Allow ordering or automatic payment within 15 days. The money is transferred to the sellers when the business is successful. Otherwise, the money will be refunded to the buyer. A third party payment service is widely used in the online marketplace, promoting online transactions by providing institutional credit [5]. Tobao uses Alipay as their escrow service to secure every transaction. When referring to a fraudulent transaction, the seller must pay the goods in advance to the consortium. Go through a third party payment company and get the money back to the seller. This process usually takes two to five days due to the time taken for delivery. This will take longer if buyers forget to verify the reception. Each product is very expensive but the time spent spending money is high. More can be traded for cheaper products for the same amount. Therefore, the fraudulent shopkeeper prefers less valuable items while improving the joint seller reputation scores. Therefore, low value items are the best choice to get a higher positive rating‌ with the same amount. Chu et al. [17] Inferior people give birth to inferior offspring and, thus, propagate their inferiority. As mentioned above, we come to the next:

Hypothesis 2 (low value production theory): Counterfeit transactions prefer cheaper products to more expensive ones. E-commerce platforms generally take steps to detect and prevent fraudulent transactions. Then they close fraudulent accounts. Collective firms should continue to register new accounts to ensure appropriate accounts are distributed to fulfill collective responsibilities as it is reasonable to conclude that many newly registered accounts are involved in fraudulent business. Therefore, we construct the following concept:

Hypothesis 3 (New Registered Fraud Buyers Theory): Newly registered accounts are more likely to participate in collective transactions. Counterfeit transaction groups try to avoid risks found by the website operator, who are required to provide counterfeit express logistics information even if they do not deliver anything. If the transactions indicate explicit products, illegal companies are required to purchase express numbers from express companies for fake express information. Therefore, the cost of counterfeit delivery will increase. Cleverly, collective societies are reluctant to deal with fees. They must take into account the nature of the product. The practical solution is to transact a virtual product that does not require physical delivery. They are: online software, music, digital photos, game card, telephone recharging, etc. Therefore, the following theory is put forward:

Hypothesis 4 (friction product nature hypothesis): The collection company prefers virtual products to explicit ones.

# 2.6 Conclusion

In our experience, we use the real-world dataset provided by Tobao Platform, one of the world's leading e-commerce websites [4]. The platform was established in 2003. It took two years for Tobago to top the China e-business list. Our data was obtained under an unknown agreement. The data includes fraudulent accounts and transaction documents related to fraudulent transactions. Several attempts were made to find a collective bargain. There are two types of detection systems. One is the mode that works on the complaint, the other is the all-selling trial. The mode that works on the complaint indicates that Tobao is investigating the specific seller the buyer is complaining about. If the seller cannot provide strong evidence to explain the suspicious case, he will be printed as a fraud account. Reputable Tobao All sellers are investigating to find a nig company in inflation. The goal of this platform is to identify whose popularity is growing rapidly. We have two types of data. We used a self-developed web crawler program to capture bad commercial records. It contributes to the structure of our transaction network. We collect data about well-behaved customers associated with fraudulent sellers labeled using the platform. Finally, we found that most legitimate consumer information through the platform is unfavorable to users.We use the data collected from Tobao as two variables: virtual and product type. ThelabelIf Virtual is a good dummy variable, however, used to determine whether a product is a physical asset. If a product is virtual it is zero. We use variable product type to indicate product type. There are 16 categories in total. When we sort products, we refer to the 16-category list from the Tobao homepage.

In our empirical study, the dependent variable is the dummy variable. If the transaction is fraudulent, label 1, and 0 will be used to label the blank transaction. In our study, we collectively consider a transaction when the seller and the buyer are fraudulent. The unfavorable consumer on the one hand, however, on the other hand, usually receives payment from the seller when the buyer agrees with the con- companies. This law is published to protect the interests of the seller if the buyer forgets to verify the order. The system predicts trade well by default.

# Chapter 3: Methodology

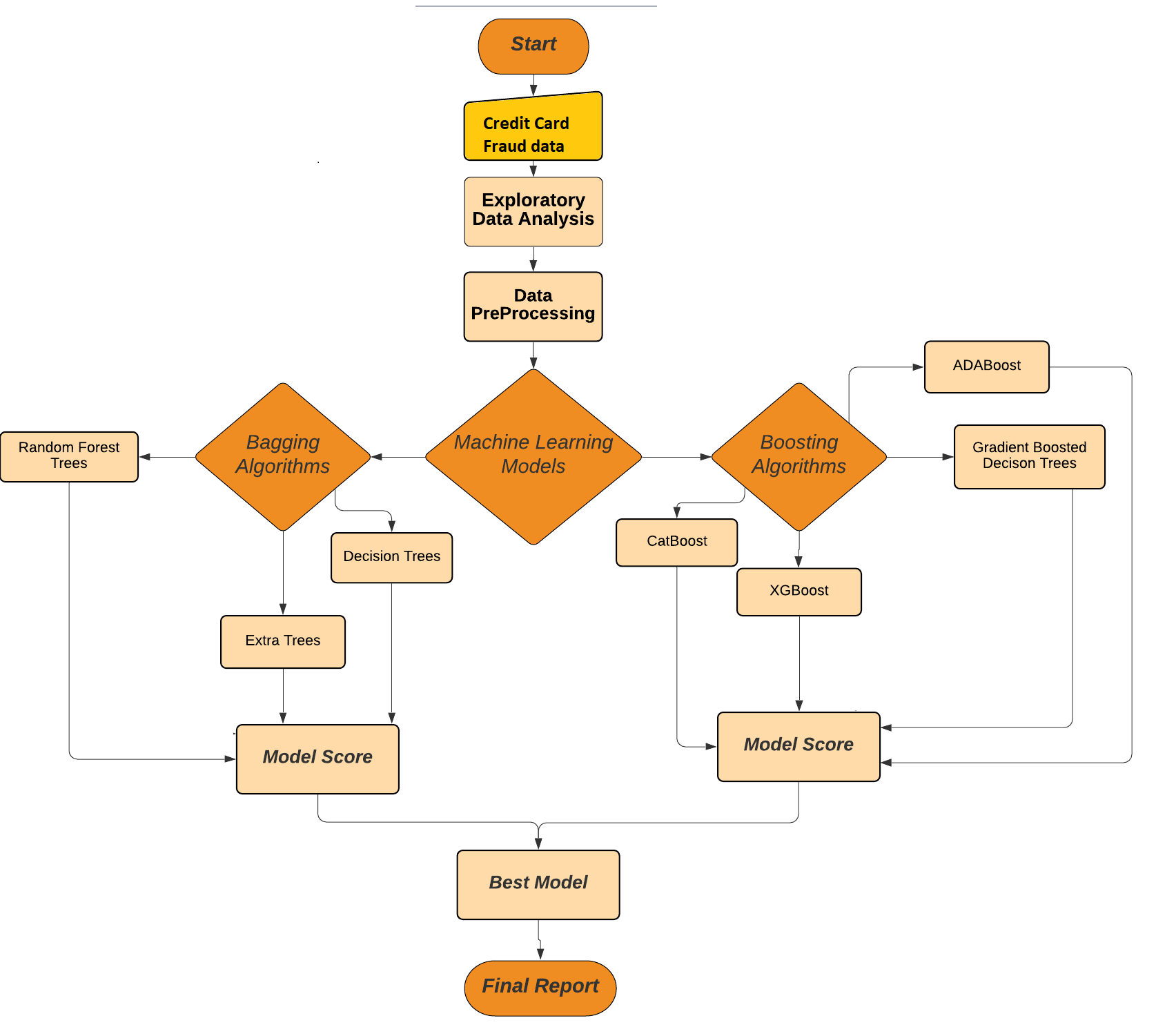


Figure 1: Design Flow

The design starts by importing the credit card fraud data available on https://data.world/vlad/credit-card-fraud-detection into the python notebook environment in local machine. The libraries of python that are required for all the data processing model implementation and evaluation are imported into python notebook.

Then the task of performed an exploratory data analysis. This step it is checked for its structure, null/missing values, class balancing, and outliers. All these issues are treated in the step of data pre-processing, where data is processed to be clean and ready for the execution in the machine learning algorithms and encoding the data for the classification models. The first task in the data pre-processing step is to remove the missing/null values. This can be done in two ways by either removing the missing values or by adding a default value.

Then the task performed next is feature selection. This is an important step as the accuracy of the model depends a lot on the features that it is built on. The overfitting/underfitting of the models, which means in overfitting running the machine learning with too many features and some of them are insignificant. Underfitting is training the model with too few features that it is not able to perform with the right accuracy. This is treated by checking the features for their level of significance and the features that are best for training the model will be the only features kept in the data.

The classifications models work best on categorical data in form of numeric labels rather than string labels. Then encoding on the data is done, to convert such features so the model efficiency is improved. Encoding the data also helps in the fast processing of the models, when the data is huge model training part can run up to days.

Then the implementation of all the five models proposed namely: CatBoost, ADABoost, Gradient Boosted Decision Tree, Light GBM and XGBoost as boosting models, and Bagging Decision Tree, Random Forest Trees and Extra Tree as Bagging Models. For every algorithm, saving the accuracy score, F1 score, precision and recall in a separate table so comparative analysis can be done effortlessly. The predictions made by the models are also compared with the actual values. The output result of each step is discussed in the next section of the finding analysis.

# Chapter 4: Results and Analysis

# 4.1 Data Exploration

The data set chosen for data exploration is as below figure and it has numerous columns, the preprocessing and exploration of data would remove columns that are not significant for our analysis and process. The same can be compared between figure 4.1 and figure 4.6.

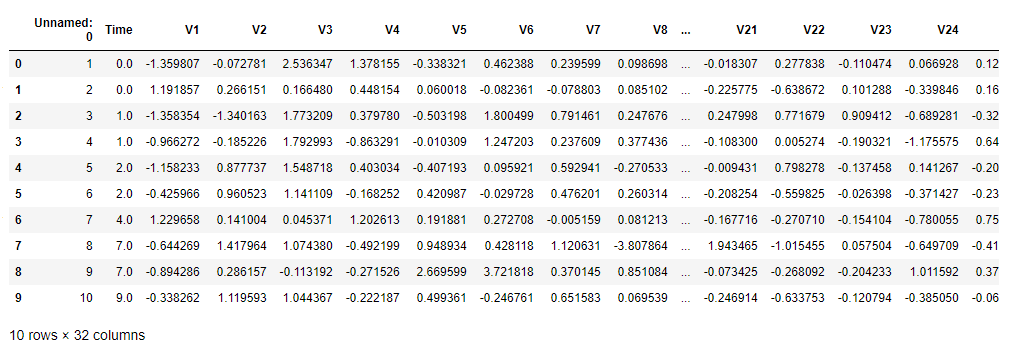


Fig 4.1 Sample data Screenshot

The below figure 4.2 samples the data and identifies the schema of the dataset and the schema identification is completed, and results are displayed as below. Note all the columns except first and the last columns are all float and excluding one date column. There are a total of 32 columns and the range of the index of the dataset and the space the dataset occupies are also listed in the below screengrab.

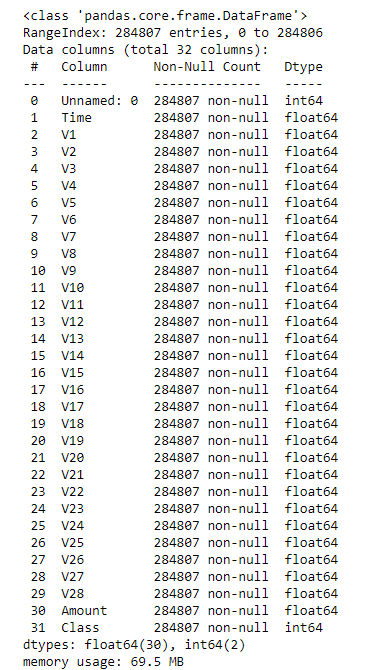


Fig 4.2 Initial Dataset Schema

The below figure shows the class distribution of the dataset, the classes are distributed among one and zero and from the below figure it is clear that the class that is represented with skewness is majority class 0, with minority class 1. This analysis helps to further understand the data under discussion, and it is agreed that the data has to be re classified so the distribution of data is even and without skewness. The process would involve reformatting the data and then to remove data rows such that the skewness issue is resolved.

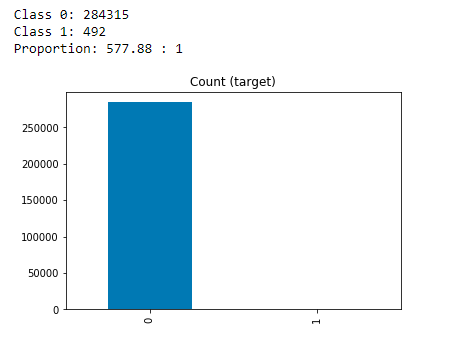


Fig 4.3 Class proportion of initial data.

After analyzing the class distribution in the above diagrams, heat diagram representation of the data under question is represented in figure 4.4, this visualization shows the distribution of data between the columns and columns that have over representation and under representation in the data set. By choosing a column that is represented properly the final result would be accurate and up to the point.

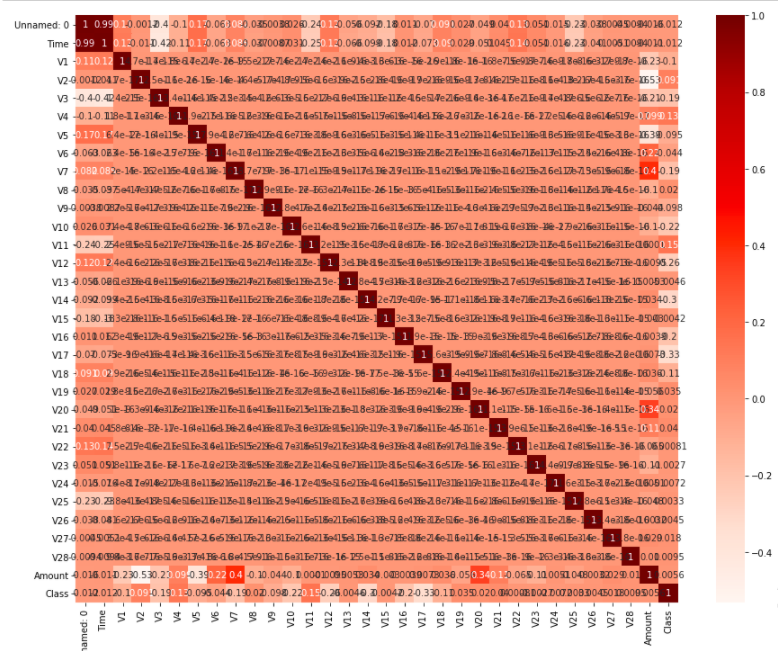


Fig 4.4 Data heat map

After analyzing the heat map and class distribution of the data set the data is cleansed and all attributes that are insignificant to the result are removed and cleaned. And the final output of the data is represented in the class distribution diagram shown in figure 4.5, here the distribution of classes one and two are almost evenly distributed for further processing. The columns that are significant and has been kept for further analysis are discussed in the next figure 4.6.

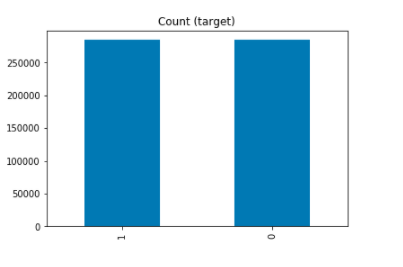


Fig 4.5 Data distribution after class Balancing

The below diagram Figure 4.6 shows the final list of columns that are selected for further processing and one can note the final list of columns is halved to 15 count compared to the initial 32 count of columns in the original data. Also the data is distributed properly between classes to avoid skewness.

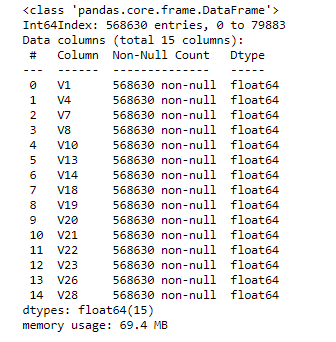


Fig 4.6 Final Data after class balancing and removing in significant attributes.

# 4.2 Data Analysis

## 4.2.1 What is Done to the Data

The initial data is of 32 columns and heavily skewed towards one class of records, any results obtained from the said data would also be skewed and would not return accurate and acceptable results. For making the data proper for the models to ingest and produce results the data has to be freed from skewness of one particular type of record and the data has to be cleaned of unwanted columns that do not alter the final output of the results. The is well discussed in Figure 4.1 and 4.2 in the above sections.

## 4.2.2 Implementation of data Cleaning and Analysis

The skewed data thus obtained is cleaned of the said irregularities and unwanted information as discussed in section 4.2.1 and the final output of the cleaned data has got class distribution evenly distributed and only contains columns that would be of any significant to the models. The final list of columns are 16 compared to the initial list of 32 columns.

Random over sampling method is used to remove the skewness across the dataset. And the resultant data has data distribution as show in below figure 4.2.1.

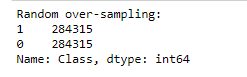


Fig 4.2.1 data representation after removing skewness across data set between record types.

Then the data is analyzed using significance tests, so the data columns that are in significant to the result would be removed from the dataset. Significance test and data preparation are done to the dataset and the columns are removed that are in significant.

## 4.2.3 result

The result thus obtained from the above processed are well formatted and insignificant columns removed and class distribution corrected to avoid skewness.

The data thus prepared is represented in below figure 4.2.3.1.

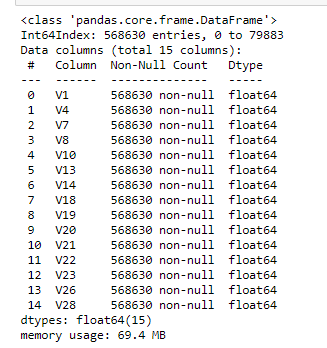


Fig 4.2.3.1 Final list of columns and index count of the processed data.

# 4.3. Machine Learning

In this section the various Models of supervised and unsupervised learning are done and the same are analyzed for accuracy of results and various other metrics.

Below models are discussed:

* Gradient Boost Tree Classifier
* CATBoost Tree Classifier
* XGBoost Classifier
* ADABoost Classifier
* Random Forest Classifier
* Extra Tree Classifier
* Decision Tree Classifier

## 4.3.1 Random Forest Trees

The Random forest is a group of tree data structure that are grouped together in an ensemble. Each individual tree generates a class prediction and the class with the most votes becomes the result of the random forest tree model. The random forest tree behaves in a wisdom of crown concept.

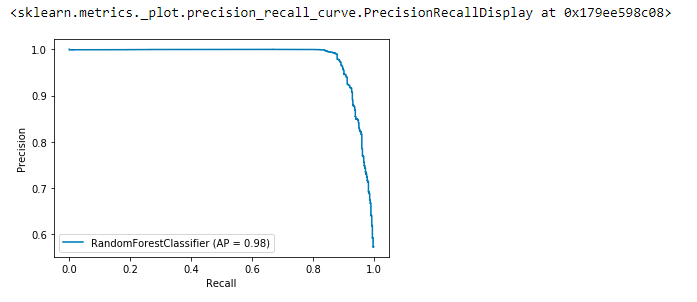


Fig 4.3.1. Random Forest Precision Recall Display

The Precision recall curve shows the comparison tradeoff between precision and recall of different thresholds. The above figure shows high area covered by the scores thereby implying that the high precision rates to low false positive rates and the high recall rates to the low false negative rate. Which means the classifier is showing accurate results.

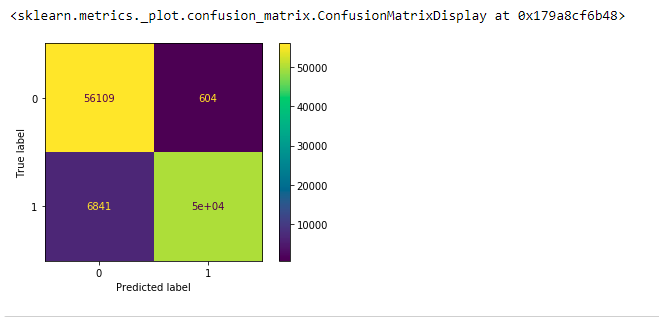


Fig 4.3.2. Random Forest Confusion Matrix

The above confusion matrix shows the comparison between accurate results and inaccurate results. As we can see the actual label is predicted most of the time and only 604 instances of 0 was predicted as 1. And 6841 instances of 1 has been predicted as 0.

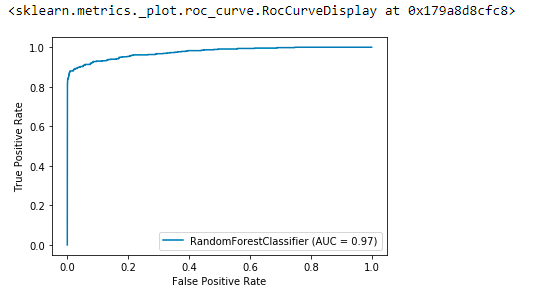


Fig 4.3.3. Random Forest ROC Curve Display

ROC curves display true positive on the X axis and the false positive on the Y axis. As it is displayed on the above figure the false positive rate is overwhelmed by the true positive rates. So the model we have generated is having high accuracy in predicting the result.

## 4.3.2. Extra Tree Classifier

This construct is very similar to random forest tree classifier the only difference is the way the construction of the decision trees. The output of Extra tree classifier would be multiple de corelated decision trees.

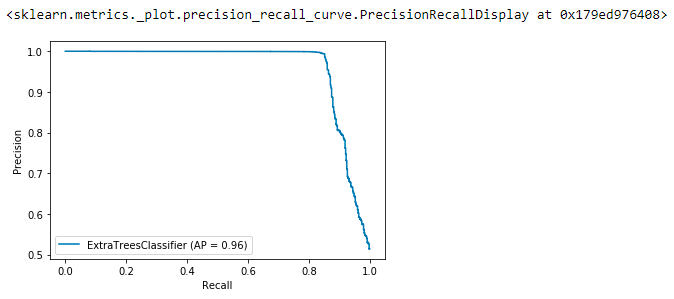


Fig 4.3.4. Extra Tree Classifier Precision Recall Display

The precision recall is displayed in the above figure. For this implementation, the curve has got high area under it, as it was the case with Random tree classifiers. So the output of the tree model is accurate with AP of 0.96.

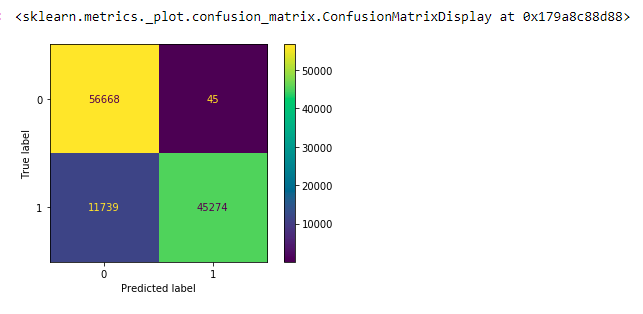


Fig 4.3.5. Extra Tree Classifier Confusion Matrix Display

From the above confusion matrix, we can see the 0 class were identified properly as compared to 1 class. Where we can see the zero class has got only 45 wrong predictions as compared to a significant amount of 1 class being classified as 0 in the model.

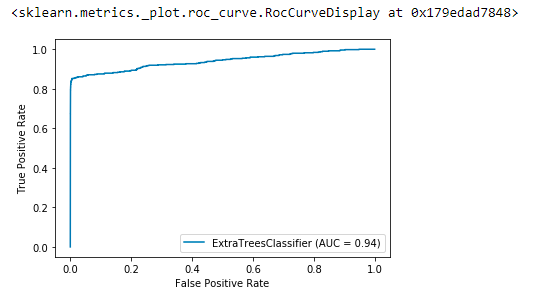


Fig 4.3.6. Extra Tree Classifier ROC curve display

The steepness of the curve shows the false positive rates are lower compared to the true positive cases. And the area under the graph represents the majority of the identifications are true positive with AUC as 0.94.

## 4.3.3 ADABoost

Boosting creates strong classifier from number of weak attributes. Ada boost is a form of boosting technique that is first adopted into solving practices.

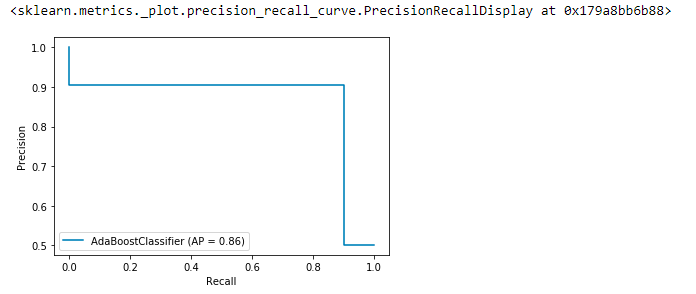


Fig 4.3.7. ADA Boost Precision Recall Display

The Ada boost precision and recall graph is not as impressive as the tree classifiers we has discussed in the previous topics. But it is having a AP score of 0.86.

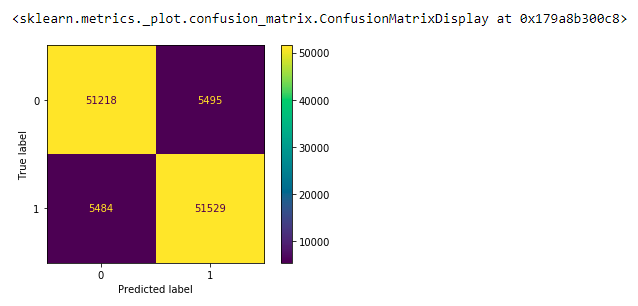


Fig 4.3.8. ADA Boost Confusion Matrix Display

The Ada Boost Confusion matrix shows the identification between 0 and 1 classifiers as they are the same and there are no differences between the actual and predicted values in comparison from tree classifiers.

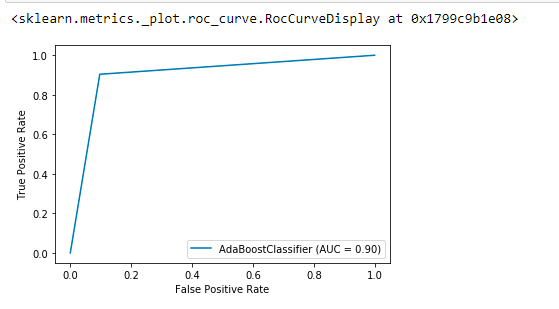


Fig 4.3.9. ADA Boost ROC Curve Display

The false positive rate is a bit higher in this model as compared to our tree models but the area under the graph is good in providing majority of results as good results.

## 4.3.4. XG Boost

XG boost is an implementation of gradient boosted decision trees. And can be used for performance and speed of execution.

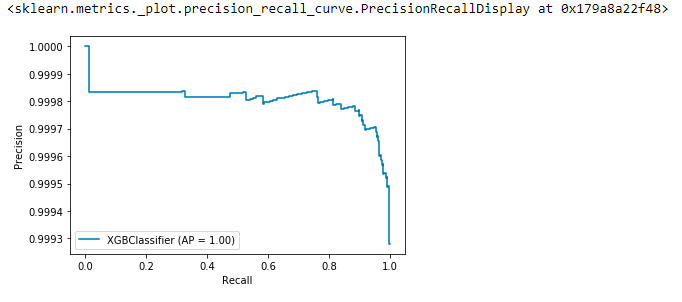


Fig 4.3.10. XG Boost Precision Recall Display

The precision recall graph shows the precision does not go below 0.9993 which is a great score and the AP score of the same is 1. So, the result obtained by XG boost is very impressive.

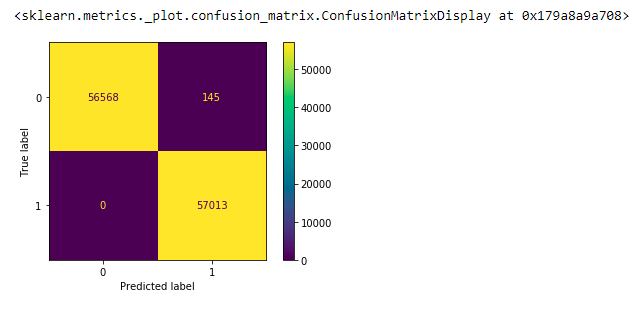


Fig 4.3.11. XG Boost Confusion Matrix

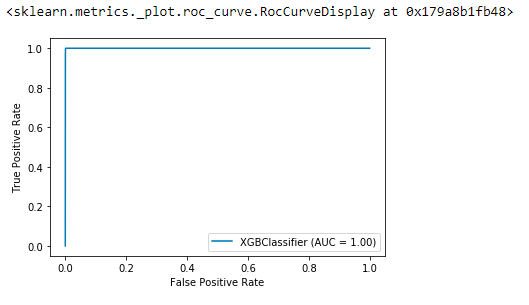


Fig 4.3.12. XG Boost ROC curve display

## 4.3.5 Gradient Boost

Gradient boost can overfit the training dataset and applies greedy approach. This increases the performance of the algorithm by overfitting.

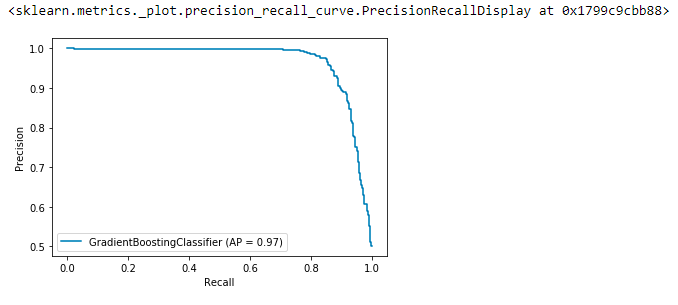


Fig 4.3.13. Gradient Boost Precision Recall Display

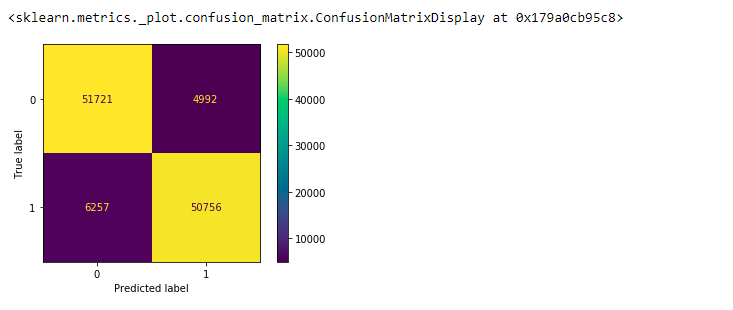


Fig 4.3.14. Gradient Boost Confusion Matrix

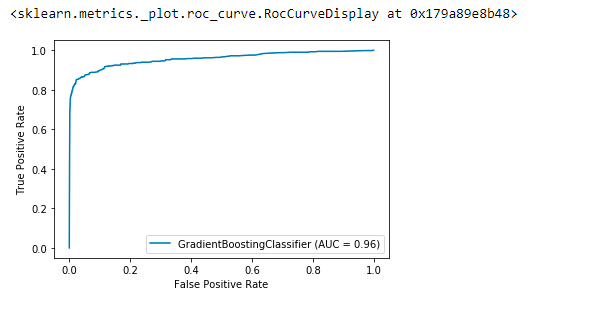


Fig 4.3.15. Gradient Boost ROC curve Display

## 4.3.6. CAT Boost

This boosting algorithm is open source developed by Yandex. Uses permutation driven alternative to solve categorical features and is at best a gradient boosting algorithm.

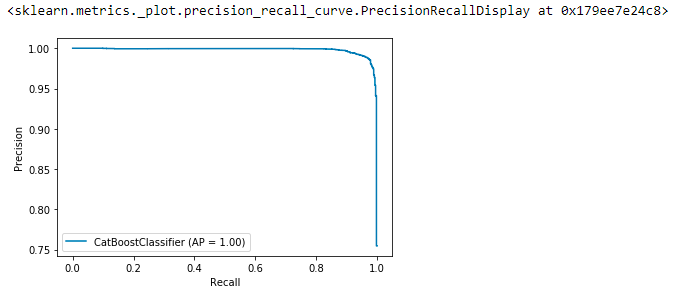


Fig 4.3.16. CAT Boost Precision Recall Display

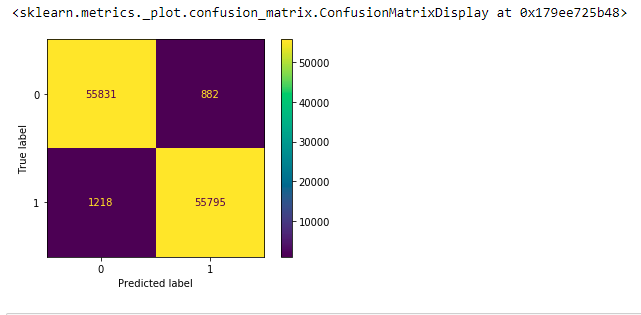


Fig 4.3.16. CAT Boost Confusion Matrix

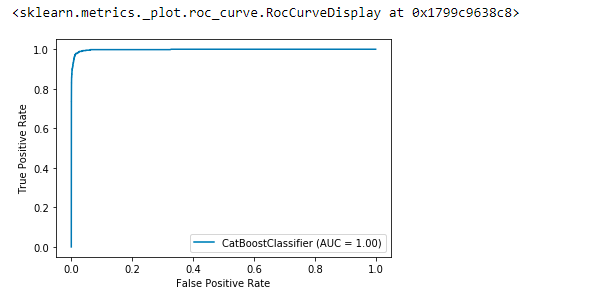


Fig 4.3.17. CAT Boost ROC Curve Display

# 4.4. Comparative analysis

The Below comparative analysis Figure shows the various Accuracy, F1 Score, Precision and Recall scores. And it is cleat that the accuracy of XG boost is best compared to other models and the overall scores of all the attributes favors XG boost for predicting in this dataset.

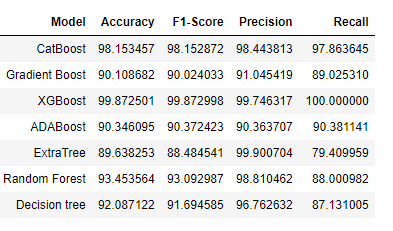


Fig 4.4.1 Model Vs (Accuracy + F1 Score + Precision + recall)

The above table shows the comparison between all the models for their attributes. And the best performer and the least performer can be identified.

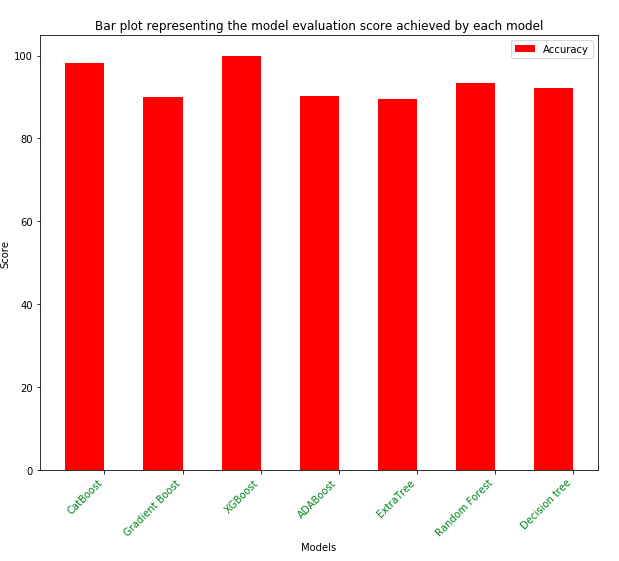


Fig 4.4.2 Model Vs Evaluation Score

The Accuracy is shown in the above graph in Red. And the XG boost tops the accuracy charts and the ADA boost and Extra tree trail in the bottom end of the comparison.

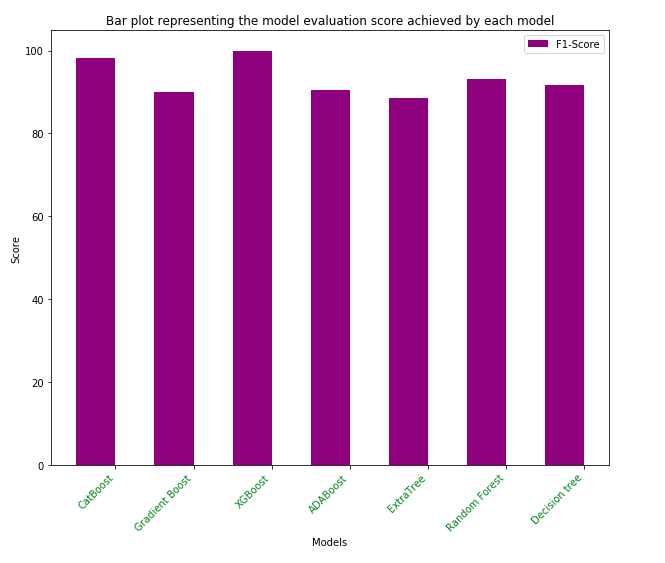


Fig 4.4.3 Model Vs F1 Score

The above purple graph figure shows the F1 scores of all the models that were evaluated and the best performer are CatBost and XG Boost. The least performer is the Extra tree.

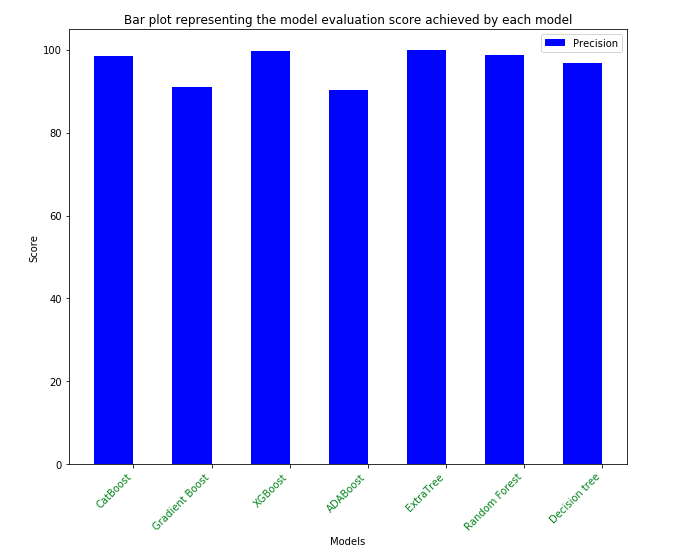


Fig 4.4.4 Model Vs Precision

The above graph shows the precision with which the detection was done with all the models generated. The best precision is provided by XG boost and the least precession is provided by ADA boost.

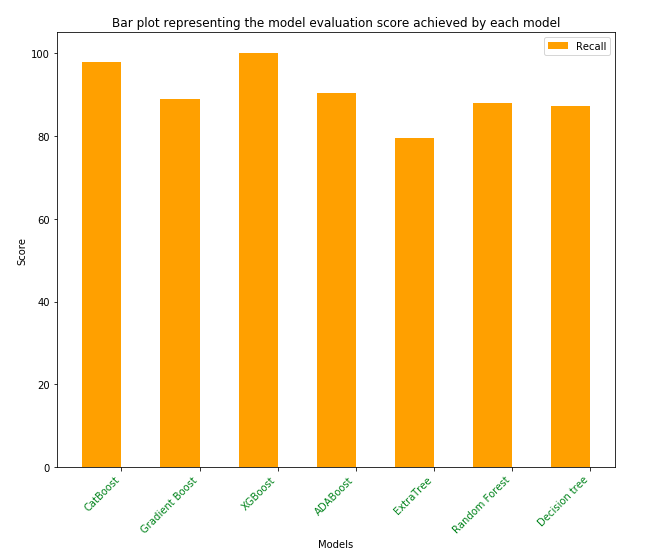


Fig 4.4.5 Model Vs Recall

The above graph in orange shows the Recall scores of the various models in question. The recall score of the XG boost is best in the series and the Extra Tree is the worst performer in the lot.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Actual** | **Cat Boost** | **Gradient Boost** | **XG Boost** | **Ada Boost** | **Extra Trees** | **Random Forest** | **Decision Tree** |
| **105178** | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| **123270** | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| **156988** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| **89432** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **247673** | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 |
| **57686** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **42856** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| **281362** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **190368** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| **27738** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| **106212** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **207092** | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| **100623** | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| **100515** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **92777** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| **45732** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| **226531** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **155634** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **40647** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **251904** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| **12108** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| **219025** | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| **143335** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| **8619** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **183078** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **6862** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **6446** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| **18792** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **15204** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| **256219** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **275992** | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| **74794** | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 |
| **127330** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **208060** | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 1 |
| **230476** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| **27362** | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| **93336** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **245377** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **136593** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **102444** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| **249014** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **95150** | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 |
| **42769** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| **275740** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **64329** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| **280916** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **221377** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **157855** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **27362** | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| **238366** | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

Table 4.4.1 Actual Vs predicted

The above table shows the actual vs predicted values for a list of sample records in the sample dataset used.

# Chapter 5: Conclusions and Future work

# 5.1. Conclusion

To summarize what we have done in this project,

* We have collected data to be used in analysis and feature creation.
* We used pandas to calculate many model features and produce pure data to assist us in machine learning. Created predictions or true values ​​using pandas.
* Cross-validation was used to avoid look-forward bias. Many machine learning models are trained to create high estimate accuracy and then integrated using synchronization learning.
* We mainly used the following mentioned machine learning model of supervised and semi-supervised category:

1. Gradient Boost Tree Classifier
2. CATBoost Tree Classifier
3. XGBoost Classifier
4. ADABoost Classifier
5. Random Forest Classifier
6. Extra Tree Classifier
7. Decision Tree Classifier

# 5.2. Future Work

Future work would use deep learning and reinforcement learning to further enhance the accuracy and predictive analysis of the model. The results from all the three reinforcement learning techniques is measures and best suited model is developed by examining the result. The three techniques which can be included in future work.

* Q-learning
* Deep Q-learning
* Dualling network

Along with these reinforcement learning techniques we can also use deep learning and Q-learning. The neural net is trained with an aid of arbitrary sample size chosen at the end of every episode in real-time. Hence after every episode the neural works in cyclic manner to collect and train data further. As a result of this cyclic process the function converges with increasing number of iterations and thus increasing the performance of the model over a period. The deep learning and neural network can be used for future work for the similar predictive analysis of alcohol detection.

**References**

1. Bharathi R (2020). Leveraging Product Characteristics for Online Collusive Detection in Big Data Transactions. *International Journal of Engineering Research and*, V9 (08).
2. Luo, S. and Wan, S. (2019). Leveraging Product Characteristics for Online Collusive Detection in Big Data Transactions. *IEEE Access*, 7, pp.40154–40164.
3. Abdul Razak, T. and Najeeb Ahmed, G. (2015). Detecting Credit Card Fraud using Data Mining Techniques - Meta-Learning. *Indian Journal of Science and Technology*, 8(28).
4. Bhattacharya, S. (2007). Review of “Essential Software Architecture by Ian Gorton,” Springer-Verlag New York Inc., Secaucus, NJ, 2006, $59.95, ISBN: 3540287132. *Queue*, 5(2), p.56.
5. You, W., Liu, L., Xia, M. and Lv, C. (2011). Reputation inflation detection in a Chinese C2C market. *Electronic Commerce Research and Applications*, 10(5), pp.510–519.
6. Wu, F., Li, H.-H. and Kuo, Y.-H. (2011). Reputation evaluation for choosing a trustworthy counterparty in C2C e-commerce. *Electronic Commerce Research and Applications*, 10(4), pp.428–436.
7. Huang, S. (2011). Designing utility-based recommender systems for e-commerce: Evaluation of preference-elicitation methods. *Electronic Commerce Research and Applications*, 10(4), pp.398–407.
8. ‌Zhang, J. (2011). Extensive Experimental Validation of a Personalized Approach for Coping with Unfair Ratings in Reputation Systems. *Journal of theoretical and applied electronic commerce research*, 6(3), pp.9–10.
9. Ying, L. (2013). A Path Analysis of Influence of Chinese C2C E-Commerce Customer Value on Customer Loyalty. *SSRN Electronic Journal*.
10. ‌Yu, Y. and Liu, Z. (2019). Multi-dimensionality reputation evaluation model for C2C E-commerce in hesitant triangular fuzzy setting. *Journal of Intelligent & Fuzzy Systems*, 37(2), pp.1809–1817.
11. Zhang, X., Luo, J. and Li, Q. (2012). Do different reputation systems provide consistent signals of seller quality: a canonical correlation investigation of Chinese C2C marketplaces. *Electronic Markets*, 22(3), pp.155–168.
12. ‌ZHONG, Y., XING, Q., LI, R., ZHANG, J. and CAO, R. (2013). Chinese Paper-cut Cultural Landscape Pattern in the C2C E-commerce Market. *Journal of Geo-information Science*, 15(4), p.560.
13. ‌Kumar Sharma, N., Gaur, V. and Bedi, P. (2016). Safeguarding Buyers with Attack-Resilient Reputation Parameters. *Journal of theoretical and applied electronic commerce research*, 11(1), pp.4–4.
14. Zhang, H., Lu, Y., Shi, X., Tang, Z. and Zhao, Z. (2012). Mood and social presence on consumer purchase behaviour in C2C E-commerce in Chinese culture. *Electronic Markets*, 22(3), pp.143–154.
15. Ng, I.C.L. and Tseng, L.-M. (2008). Learning to be Sociable: The Evolution of Homo Economicus. *American Journal of Economics and Sociology*, 67(2), pp.265–286.
16. Interdisciplinary Studies of Consciousness: From Homo Economicus to Homo Cognitivus. (2007). *Foresight-Russia*, 1(4), pp.32–35.
17. ‌Machalek, R. (2010). The Origins of Sociable Life: Evolution After Science Studies. *Contemporary Sociology: A Journal of Reviews*, 39(3), pp.306–307.
18. Pavlou, P.A. and Gefen, D. (2004). Building Effective Online Marketplaces with Institution-Based Trust. *Information Systems Research*, 15(1), pp.37–59.
19. ‌Patnasingam, P., Gefen, D. and Pavlou, P.A. (2005). The Role of Facilitating Conditions and Institutional Trust in Electronic Marketplaces. *Journal of Electronic Commerce in Organizations*, [online] 3(3), pp.69–82. Available at: <https://pdfs.semanticscholar.org/d571/6ef33440fc3f9c98041a1a11ed26d08241d4.pdf>.
20. Pavlou, P.A. and Gefen, D. (2005). Psychological Contract Violation in Online Marketplaces: Antecedents, Consequences, and Moderating Role. *Information Systems Research*, 16(4), pp.372–399.
21. Warkentin, M., Gefen, D., Pavlou, P.A. and Rose, G.M. (2002). Encouraging Citizen Adoption of e-Government by Building Trust. *Electronic Markets*, 12(3), pp.157–162.
22. ‌Pavlou, P.A. and Dimoka, A. (2006). Institutional Feedback Technologies in Online Marketplaces: An Investigation of Feedback Text Comments, Trust, and Price Premiums. *SSRN Electronic Journal*.
23. Tran, T. (2009). Protecting buying agents in e-marketplaces by direct experience trust modelling. *Knowledge and Information Systems*, 22(1), pp.65–100.
24. ‌Hong, I.B. (2018). Building Initial Trust in an Intermediary in B2C Online Marketplaces. *Journal of Global Information Management*, 26(2), pp.27–47.
25. Rana, O.F. and Kephart, J.O. (2006). Building Effective Multivendor Autonomic Computing Systems. *IEEE Distributed Systems Online*, 7(9), pp.3–3.
26. ‌Randels Jr., G.D. (1998). *Journal of Business Ethics*, 17(12), pp.1299–1310.
27. Randels, G.D. (2001). Loyalty, Corporations, and Community. *Business Ethics Quarterly*, 11(1), pp.27–39.
28. ‌Betz, J. (1998). Business Ethics and Politics. *Business Ethics Quarterly*, 8(4), p.693.
29. Hartman, E.M. (1998). The Role of Character in Business Ethics. *Business Ethics Quarterly*, 8(3), pp.547–559.
30. ‌Koehn, D. (1998). Virtue Ethics, the Firm, and Moral Psychology. *Business Ethics Quarterly*, 8(3), p.497.
31. Danley, J.R. (1991). Polestar refined: Business ethics and political economy. *Journal of Business Ethics*, 10(12), pp.915–933.
32. ‌Li, W., Wu, D. and Xu, H. (2008). Reputation in China’s online auction market: Evidence from Taobao.com. *Frontiers of Business Research in China*, 2(3), pp.323–338.
33. Li, X. and Zhang, B. (2008). Stock market behavior and investor sentiment: Evidence from China. *Frontiers of Business Research in China*, 2(2), pp.277–282.
34. ‌Tong, W.H.S. and Wong, M.B.T. (2020). Does reputation of sponsors matter in IPO? Evidence from Hong Kong. *Frontiers of Business Research in China*, 14(1).
35. ‌Li, D., Li, J. and Lin, Z. (2008). Online consumer-to-consumer market in China – A comparative study of Taobao and eBay. *Electronic Commerce Research and Applications*, 7(1), pp.55–67.
36. ‌Liu, W. and Liu, X. (2008). Auditor switching, earnings manipulation and auditor independence: Evidence from A-share listed companies in China. *Frontiers of Business Research in China*, 2(2), pp.283–302.
37. Ye, Q., Li, Y., Kiang, M. and Wu, W. (2009). The Impact of Seller Reputation on the Performance of online sales. *ACM SIGMIS Database: the DATABASE for Advances in Information Systems*, 40(1), pp.12–19.

‌