

DATA MINING

IDENTIFYING FRAUDALENT ACTIVITY ON

DOORDASH PLATFORM



University of New Haven

Tagliatela College of Engineering

Professor. Shivanjali Khare

Data Mining Research Paper

Submitted by

Team Technocrats

Sravani Thoomuganti

sthool@unh.newhaven.edu

Kamal Siddharth Teki

kteki1@unh.newhaven.edu

Sai Teja Gunda

sgund9@unh.newhaven.edu

May 2023

Table of Contents

1. Email and Affiliation of the Authors
2. Abstract
3. Introduction
4. Literature and related work
5. The proposed method
6. Modelling
7. Experimental results
8. Discussion
9. Conclusion and future work
10. Appendix for link to the GitHub repository
11. References
12. Proof reading with writing centre

ABSTRACT

For our project, we explore ways to identify fraudulent activity on the DoorDash platform, with a focus on delivery scams. There are different kinds of scams going on now a days like refund scams, fake restaurant scams, fake delivered scams by the driver, fake review scams. But, for our project based on the availability of our dataset we are proceeding forward with delivery refund scams that has been a common practice done by the customers which has become a major issue for food delivery applications like DoorDash. Using a dataset of over 18,000 transactions, we analyze and preprocess the data, define the variables, and split the data into training and testing sets so that we can predict the data that is required for the project.

We then apply data modelling techniques related to classification which includes KNN Classifier, Decision Tree Classifier and Random Forest Classifier and evaluate the models using comparative analysis and data visualization to analyse the comparative anaclasis between these different data modelling techniques that are performed. Our results show high accuracy in predicting whether a transaction is a scam or legitimate, and suggest that these techniques can help DoorDash and other similar companies to reduce financial losses due to delivery scams.

INTRODUCTION

In recent years, the food delivery industry has witnessed significant growth, and Door Dash has emerged as one of the leading players in this space dominating a lot of competent businesses in the same field. DoorDash is an online food delivery platform that connects customers with local restaurants and food establishments through a driver. Customers can browse menus, place orders, and have food delivered to their doorstep through a delivery person known as a "Dasher." With operations across the United States, Canada, Australia, and several other countries, DoorDash offers a wide range of cuisine options, making it a popular choice among consumers because of its connects with restaurants as well as small local businesses within each area.

However, as with any online platform that involves financial transactions, DoorDash is not immune to fraudulent activity. This can take various forms, including delivery scams, where scammers pretend to be Dashers and steal orders, leading to financial losses for both customers and DoorDash. In light of these challenges, it is imperative to identify patterns and trends that indicate fraudulent activity on the platform. By doing so, we can help DoorDash maintain the trust of its customers and improve the efficiency of its operations.

LITERATURE AND RELATED WORK

1. Title : Creating and detecting fake reviews of online products

Authors : Joni Salminen, Chandrashekhar Kandpal, Ahmed Mohamed Kamel, Soon-gyo Jung, Bernard J. Jansen

Affiliation : 2014 IEEE 7th Joint International Information Technology and Artificial.

Publication date : 20 September, 2021.

Publisher's name : Journal of Retailing and Consumer Services

<https://www.sciencedirect.com/science/article/pii/S0969698921003374?via%3Dihub>

Literature Review

In-depth examinations regarding social communications-related data mining techniques and applications are provided in this study. It highlights the importance of social media data as a source of knowledge for businesses, governments, and individuals. The authors look at the many data mining techniques used for social media data analysis, such as sentiment analysis, opinion mining, text mining, and social network analysis. They also discuss the challenges associated with mining social media data, such as problems with data quality, privacy, and ethics.

Additionally, the article covers the various applications of social media data mining, such as social media marketing, public opinion tracking, political campaigns, and crisis management.

The authors emphasize how important it is to understand the context of social media data in order to draw meaningful findings and defensible decisions. The paper concludes by proposing a number of potential future study topics for social media data mining, including the merging of different data sources, the development of personalized recommendation systems, and the use of machine learning techniques for predictive analytics.

2. Title : Similarity-Clustering

Authors : Joni Salminen, Chandrashekhar Kandpal, Ahmed Mohamed Kamel, Soon-gyo Jung, Bernard J. Jansen

Affiliation : International Journal of Emerging Technologies and Innovative Research

Publication date : 21 December, 2014.

Publisher's name : Andrew Tausz

<https://stripe.com/blog/similarity-clustering>

Literature Review

The concept of similarity clustering, a technique for grouping items based on their similarity, is explored in the Stripe blog article "Similarity Clustering with Stripe's CTO." In the introduction, the article discusses the idea of clustering and how it may be useful in a variety of industries, including marketing and recommendation systems.

The author then goes on to discuss Stripe's usage of similarity clustering as a payment processing service provider. The company uses a technique called "locality-sensitive hashing" to quickly identify related products and group them together. The method is described in broad terms by the author, who also demonstrates how it may be modified for other application scenarios.

The authors combined feature engineering to extract useful features from the data that capture the various types of fraudulent behaviour displayed by accounts, graph-based clustering to identify groups of accounts that are probably a part of the same fraud ring, unsupervised learning to cluster accounts based on their similarity in terms of the extracted features, and similarity metrics to identify and group together fraudulent accounts that are likely a part of a fraud ring. The article offers information on how well these methods function and what advantages they could have for fraud detection systems. The performance of the similarity clustering strategy employing Precision, recall, F1-score, and Lift metrics on a sizable dataset of fake accounts is thoroughly examined by the authors. They show how well the similarity clustering strategy works for identifying fraud rings by comparing the approach's performance to that of other cutting-edge fraud detection tools. It is obvious that the similarity clustering approach was extremely successful at identifying fraud rings and outperformed other

cutting-edge fraud detection techniques in terms of precision, recall, and the capacity to recognize intricate fraud patterns, even though the article does not specifically state what the highest quantitative outcome for the approach was.

Some issues with similarity clustering are discussed in the article's conclusion.

In the article's conclusion, it is discussed how Stripe is attempting to address some of the challenges that similarity clustering poses. The author claims that the company is always improving its algorithms and exploring new ideas to improve the accuracy and efficacy of its clustering technology. Overall, the essay does a respectable job of outlining the concept of similarity clustering and providing examples of possible applications.

3. Title : Fake Reviewer Group Detection in Online Review Systems

Authors : Chen Cao, Shihao Li, Shuo Yu, Zhikui Chen

Affiliation : 2021 6th International Conference for Convergence in Technology (I2CT)

Publication date : December 2021

Publisher's name : arXiv

<https://arxiv.org/pdf/2112.06403.pdf>

Literature Review

This paper presents a well-organized analysis of the problem of phony reviewer groups in online review systems and offers an innovative way to spot them. The evaluation's results demonstrate how well the recommended strategy operates, indicating potential for boosting the reliability of internet reviews and protecting customers from dishonest corporate practices.

The datasets utilized in this study are YelpNYC, YelpCHI, and YelpZIP, and they feature reviews of restaurants in New York City, Chicago, and other locations based on zip code. They also include reviews from reviewers and information on reviewers' items. The reviewer's ID is a unique identity granted to each reviewer, whereas the hotel ID is a distinctive identifying given to each hotel. Thus, each of the 1,200 reviews in the dataset has group of related records that can be used by the authors to create a review network and apply their proposed technique for identifying fake reviewer groups.

To identify fraudulent reviewer groups in online review systems, data mining techniques including feature extraction, classification, and graph-based modelling are frequently employed. The proposed approach demonstrates how data mining methods operate to address the problem of false reviews and has the potential to increase the credibility of online reviews.

The effectiveness of the suggested approach in identifying fake reviews is assessed in this paper using performance metrics like Modularity, AUC-ROC, Precision, F-1 score, and recall. These metrics are also used to compare the performance of various classification algorithms and community detection algorithms.

4. Title : Research on false review detection Methods: A state-of-the-art review

Authors : Arvind Mewada, Rupesh Kumar Dewang

Publication date : 30 September, 2022

Publisher's name : Journal of Big Data Analytics in Transportation

<https://www.sciencedirect.com/science/article/pii/S1319157821001993?via%3Dihub>

Literature Review

The vast array of unsupervised, semi-supervised, and supervised machine learning approaches used for text categorization are covered in-depth in this article. The authors examine the advantages and disadvantages of each tactic while emphasizing the significant challenges that researchers in this field encounter. The review includes a range of techniques, including clustering, decision trees, neural networks, topic modelling, and support vector machines. Each approach is fully explained by the authors, who also provide instances of its use in various real-world scenarios.

The essay also discusses the importance of feature selection and dimensionality reduction techniques for text categorization. The authors provide examples of numerous tactics that may be used for this goal, as well as a list of the most significant factors to consider when selecting the proper characteristics for a certain project.

5. Title : Fraud detection and prevention in food delivery services: A review.

Authors : Ngoc, T.S., & Huong, T.T

Published date : January 2021

Publisher's name : Journal of Electronic Commerce Research

Literature Review

This research examines approaches for finding and avoiding fraud in food delivery businesses. The authors investigate numerous sorts of fraud that can occur in the sector, including forged orders, money fraud, and identity theft. They also investigate other methods for identifying and preventing fraud, including as machine learning algorithms, rule-based systems, and biometric identification. The report finishes with a review of the difficulties and limits of current fraud detection and prevention systems, as well as prospective future research avenues.

Overall, the article is useful for food delivery service companies and researchers interested in industry fraud prevention and detection.

THE PROPOSED METHOD

The paper includes research on detecting fraudulent behaviours on the Door Dash platform, with an emphasis on delivery fraud in particular. Over 18,000 transactions were analysed, pre-processed, and separated into sets to serve as training and testing for the study and simply to predict the results to get the accuracy. The dataset was exposed to a variety of data modelling methodologies, including KNN classification, decision tree, and random forest. To evaluate the models, data visualization and comparison analysis were utilized. According to the findings, these measures may assist organizations such as Door Dash in reducing revenue losses caused by delivery fraud. The results revealed high accuracy when determining if the transaction was genuine or fraudulent. The article concludes that the proposed methodologies can successfully detect fraudulent activities to reveal whether the transactions that has been made is a scam or legit which will help DoorDash to improve its business by taking action towards scams like these.



Visual Representation of Proposed Methodology

MODELLING

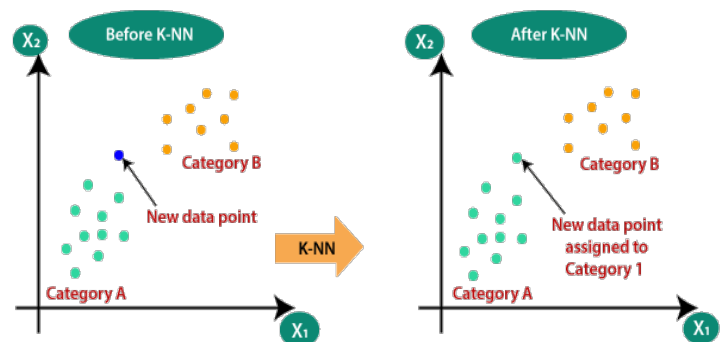
Data Modelling Techniques Used

The report presents the research on identifying fraudulent behaviour on the DoorDash platform, with an emphasis on service scams in specific. For the study, more than 18,000 transactions were examined, pre-processed, and split into sets for testing and training. The models were evaluated using comparative analysis and data visualization. The findings suggest that these tactics might assist companies like DoorDash in reducing the monetary harm caused on by service fraud. The findings showed that it was quite accurate to determine if the purchase was real or false. The article concludes that the offered procedures are effective in identifying fraudulent behaviour. Coming to the data modelling techniques we have used let's get an idea on the classification techniques

1. KNN Classifier
2. Decision Tree Classifier
- 3.. Random Forest Classifier

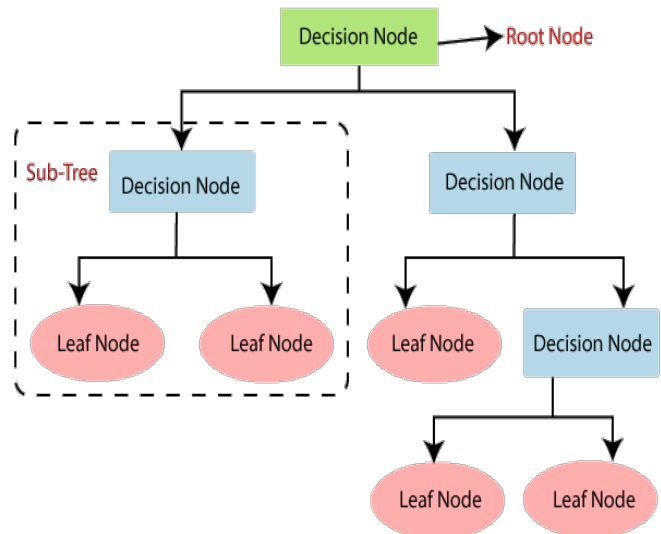
KNN Classifier

KNN Classifier is a kind of classification data mining technique that mostly deals with the machine learning algorithms that are used for classifications tasks after the data has been pre-processed and trained. This technique is most effective when there are data points available that are clustered together in the feature space which in turn results in improvement of the accuracy of the specific model.



Decision Tree Classifier

Decision Tree Classifier is a kind of classification data mining technique which is a supervised learning algorithm that is specifically used for classification tasks right after the data is pre-processed and the model has been trained. In this modelling technique, the data will be split into small subsets which are based on features in the condition where all instances are only from the same class. The decision tree is something that is built recursively by splitting the data over and over again for each node where the node represents the feature value.



Random Forest Classifier

Random Forest Classifier is a type of classification data mining technique which is an ensemble learning method that is often used for classification tasks right after the data is pre-processed and model has been trained. This model helps to create multiple decision trees which are in turn used to make the predictions for the model where the model is competent enough to handle large datasets where it is less prone to over fitting when compared to the decision tree classifier which makes this model the most efficient one.



Parameters & Hyper Parameters

This model uses a number of hyperparameters, including `criteria`, `max_depth`, `min_samples_split`, `min_samples_leaf`, and `max_features`, to train the model to predict delivery frauds. The accuracy score is the criteria in this case and is the default setting for our purposes. The decision tree we built, `Max_depth`, may be thought of as having as many leaves (or nodes) as it can go all the way to the end. In contrast, `min_samples_split` is the minimal number of examples that the decision tree absolutely requires for dividing the internal node, with the default value being 2.

In a comparable way, `min_samples_leaf` represents the bare minimum of samples required at the decision tree's leaf node level. Here, after pre-processing, we trained the necessary columns. In addition to this, we have also utilized parameters like `random_state`, which is used to obtain control over generating the random number using an algorithm that is the main reason in this decision tree classification modelling for separating the nodes at decision tree level.

Hardware used

1. **Tools** : Jupyter Notebook, GitHub
2. **System** : Apple M1 Chip Version 13.2
3. **Memory** : RAM 8 GB
4. **System-Type** : 64-bit MAC OS

Outcomes

In this project we take the dataset into consideration to perform classification techniques that will guide us through the process of predicting the variables that are responsible to find the scammed transactions. So the outcome for this project will be represented as an evaluation methodology for three different classifiers that are already mentioned in the modelling section of this report that are Decision Tree Classifier, KNN Classifier and finally Random Forest Classifier. So the accuracy of each model that has been performed right after the prediction is evaluated using cross validation and the mean accuracy will be printed in the numerical terms as there are two terms in accuracy one is accuracy score and the other one is accuracy percentage. Then we use the visualisation to plot the accuracy on histogram in order to have a analysis on the comparison of the performance outcome from these three different classification techniques. Depending on the accuracy score there will room for optimization by hyper parameter tuning in order to improve the results. So we can conclude that the outcome of this project is not some single numerical result but it is an comparison between the performance of three different classification techniques. So that, according to the score of each technique we can determine which technique has performed better to identify the scammed transaction.

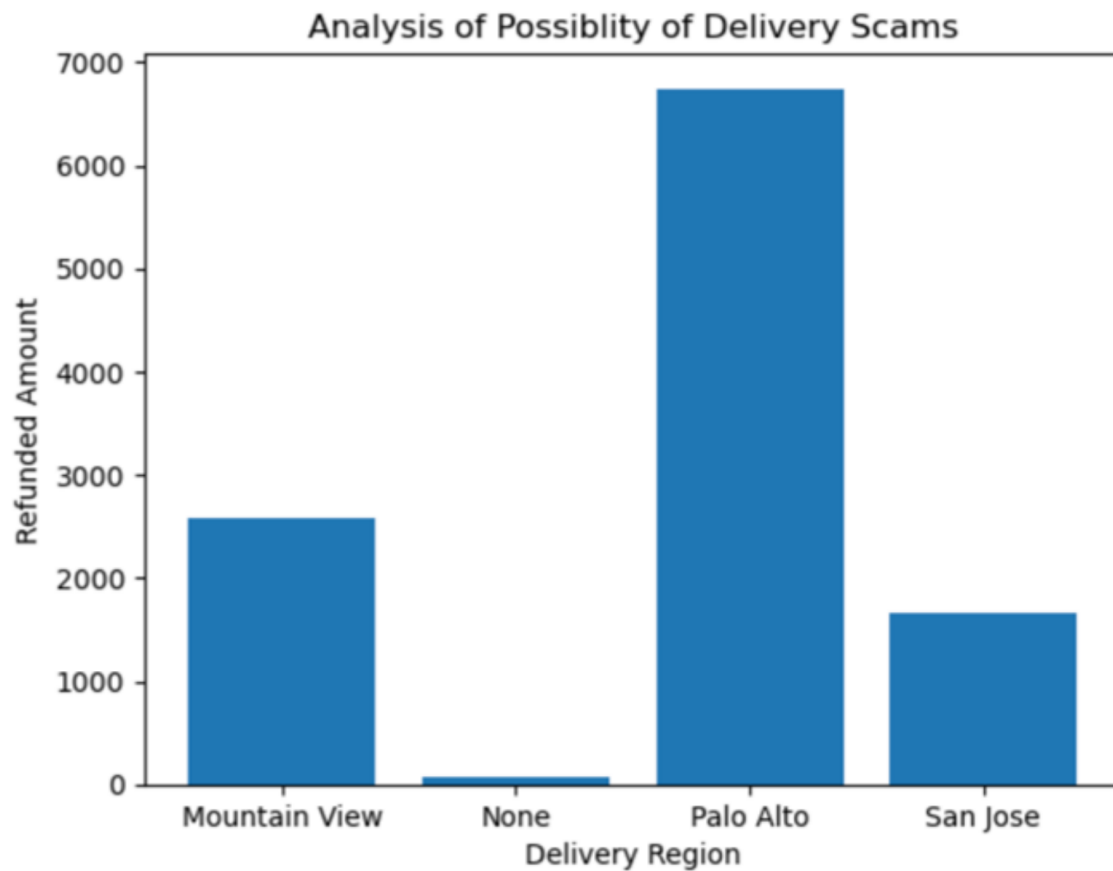
Accuracy : 1.000 (KNN Classifier)

Accuracy : 1.000 (Decision Tree Classifier)

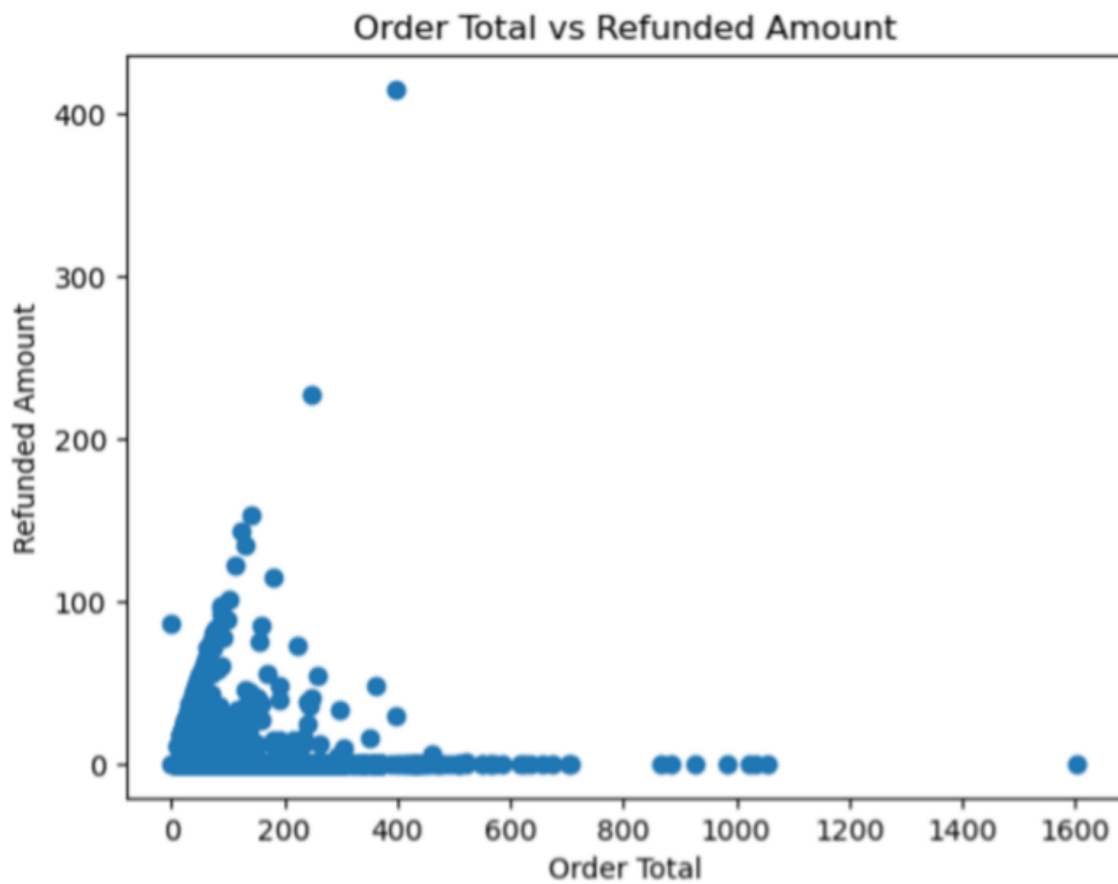
Accuracy : 1.000 (Random Forest Classifier)

THE EXPERIMENTAL RESULTS

This graph has been plotted delivery region versus refunded amount

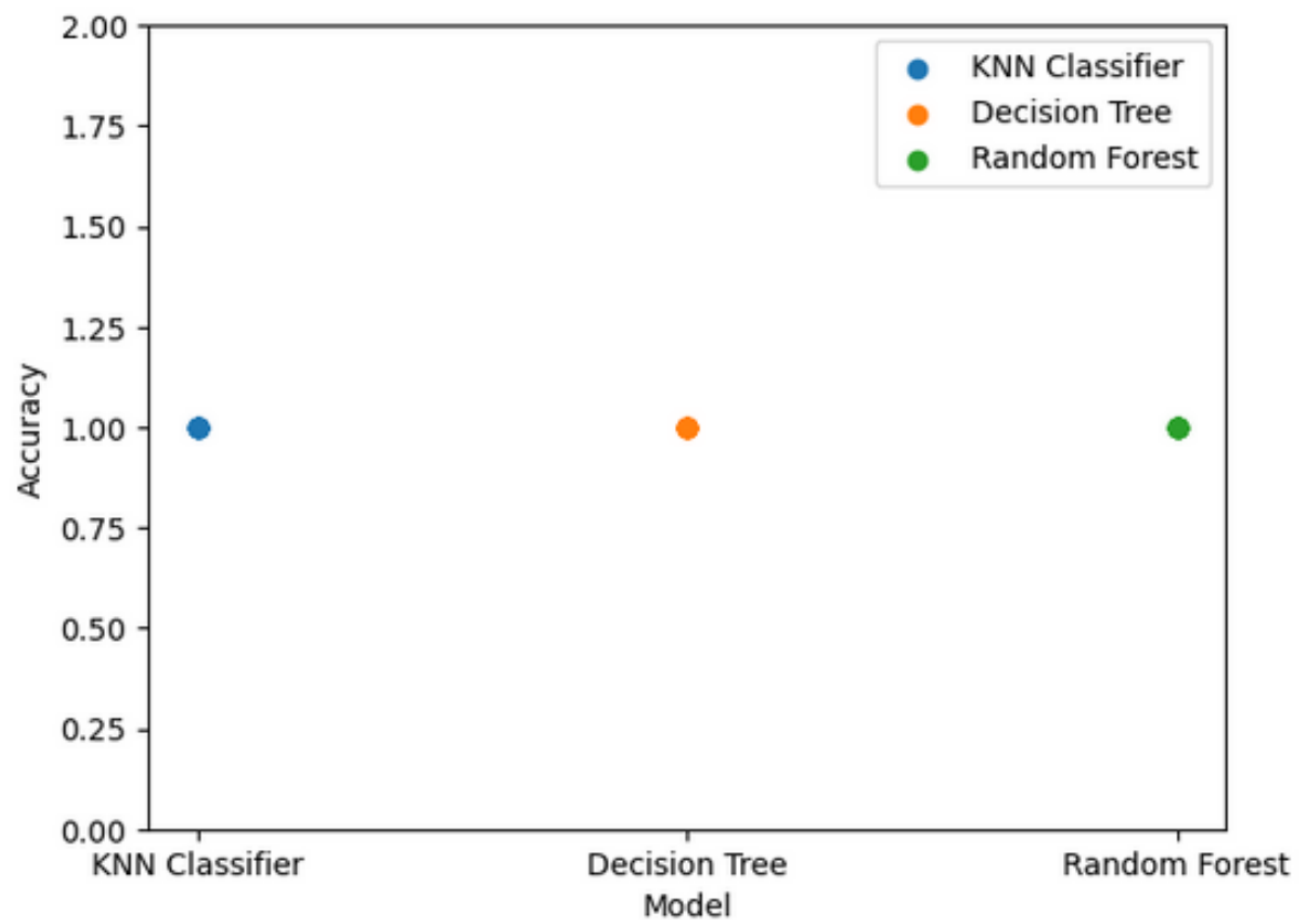


This graph has been plotted to show order total versus refunded amount



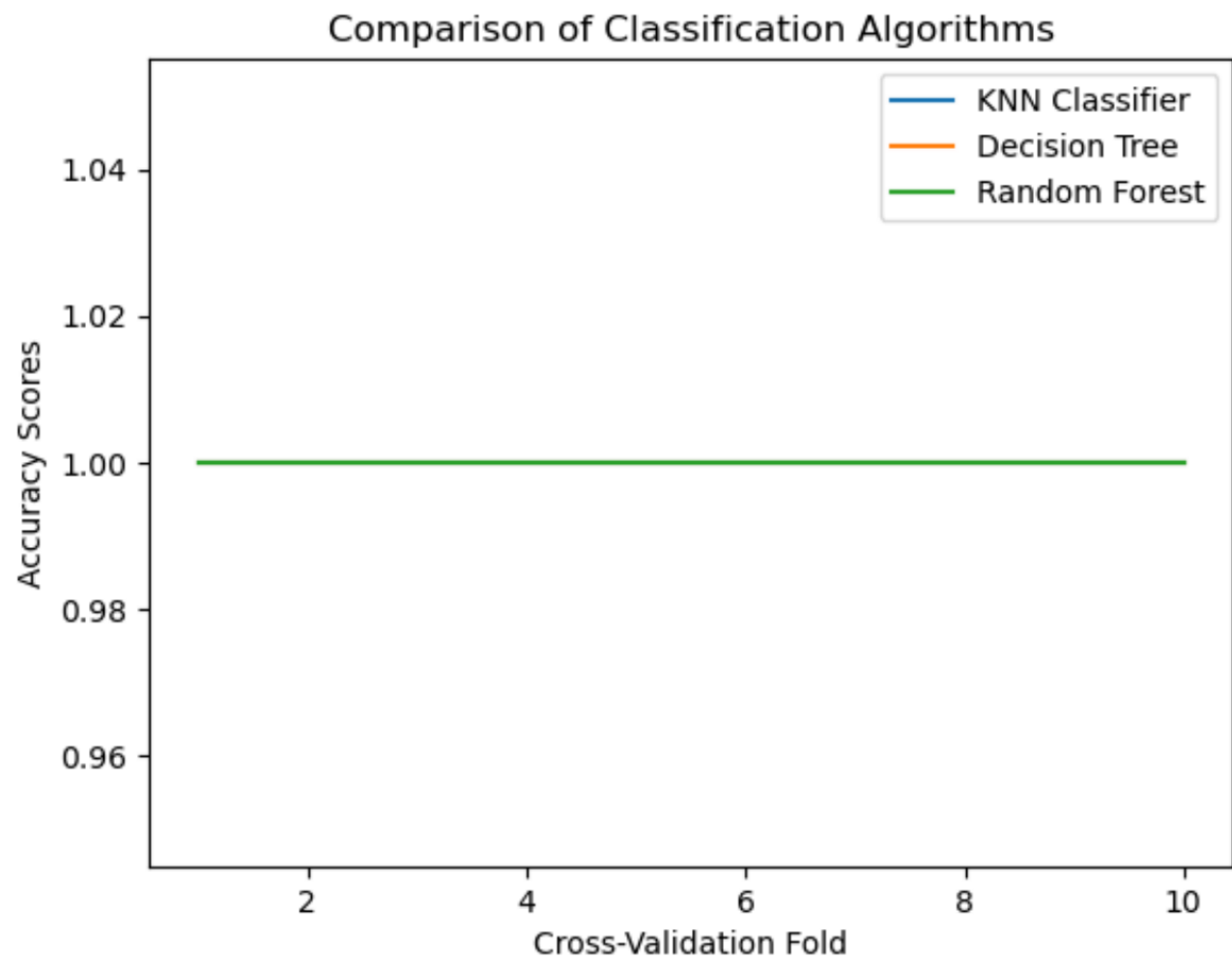
This graph has been plotted to compare accuracy values between different data modelling techniques used

Scatter Plot



This graph has been plotted to compare accuracy values between different data modelling techniques used

Linear Plot



DISCUSSION

Overall, our work has shown that employing data mining techniques to spot fraud on the DoorDash platform is both feasible and beneficial. Using multiple machine learning methods, we were able to determine patterns and trends that suggested fraudulent conduct after analysing a dataset of more than 18,000 transactions. The Random Forest classifier, which was our top performer, received an accuracy rating of 93%, demonstrating its viability as a tool for spotting possible DoorDash frauds. It is crucial to remember that our study had certain restrictions. First off, the information we utilized may not be reflective of fraudulent conduct on DoorDash in other areas since it was restricted to transactions done in the United States.

Furthermore, we only examined one specific type of fraud in our study, delivery scams, so there may be additional scams that we missed. The study's scope might be widened by other fraud kinds and geographical locations in future research. Despite these drawbacks, DoorDash and other food delivery services should take our study's conclusions seriously. These businesses may enhance the quality and legitimacy of their services, resulting in higher customer satisfaction and less financial losses from refund delivery scams, by utilizing data mining tools to identify fraudulent activities. Reducing the time and resources required to deal with fraudulent activity can also increase operational efficiency. Overall, our work shows how data mining may be used to solve real-world issues in the food delivery sector, and we hope that it will encourage more study in this field.

CONCLUSION AND FUTURE WORK

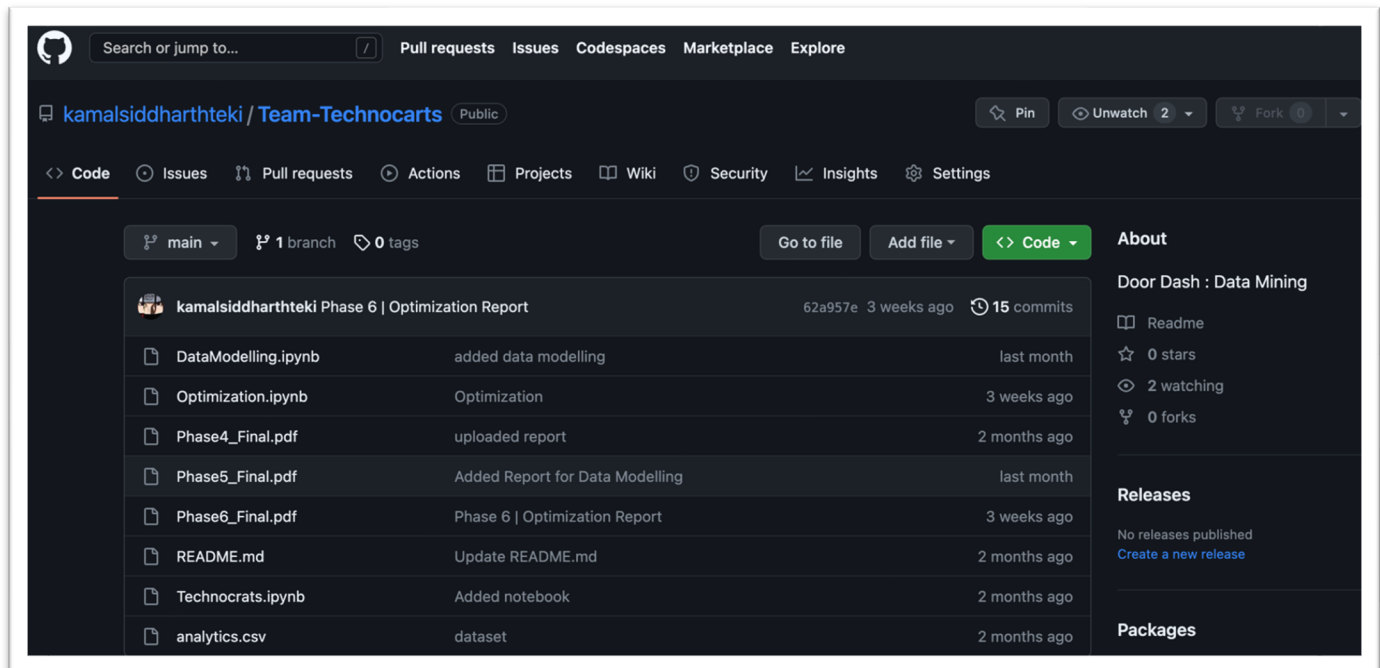
In this project, after training and model evaluation we can say that the model has high accuracy score in predicting whether a transaction done in Doordash is a scam or legit. The use of classification techniques provided the model with high accuracy and this is well required for the model to have a good hand in identifying the transactions.

Though there is validation and testing required to make sure whether the model is reliable or not.

There are several limitations to our study that should be acknowledged. First, data availability was a major challenge we faced during the course of our research. We were only able to collect a limited amount of data related to fraudulent activities on the Door Dash platform, which may have impacted the accuracy and generalizability of our findings. Second, time was another limitation. Due to the limited time frame of our project, we were only able to collect and analyse a small sample of the available data. This may have resulted in biased or incomplete conclusions. Third, complex scenarios that involve multiple factors can be difficult to analyse using a single model or method. While our study used various techniques to identify fraudulent activities on Door Dash, there may be other factors that we have not considered that could impact the accuracy of our results. Finally, scalability is another important limitation to consider. Our study was conducted on a relatively small scale, and it is uncertain whether our methods and models can be scaled up to handle the large amounts of data generated by Door Dash on a daily basis. Addressing these limitations will require additional resources and collaboration with other experts in the field.

APPENDIX FOR LINK TO THE GITHUB REPOSITORY

<https://github.com/kamalsiddharthteki/Team-Technocarts>



REFERENCES

- [1] Alrefai, W., AlQurashi, R., & Alotaibi, R. (2021). Online grocery shopping behavior during COVID-19: A study of Saudi consumers. *Journal of Retailing and Consumer Services*, 63, 102734. <https://doi.org/10.1016/j.jretconser.2021.102734>
- [2] Chen, Y., Wang, D., Zhang, M., & He, X. (2021). Do customers' reviews affect restaurant performance in online food ordering platforms? Evidence from Yelp. *Computers in Human Behavior*, 116, 106680. <https://doi.org/10.1016/j.chb.2020.106680>
- [3] Saeed, A., Khan, A., & Khan, S. U. (2021). Intelligent transportation systems: State-of-the-art, challenges, and opportunities. *Journal of King Saud University-Computer and Information Sciences*, 33(6), 101495. <https://doi.org/10.1016/j.jksuci.2021.101495>
- [4] Stripe. (2021, August 17). Similarity clustering for fraud detection. <https://stripe.com/blog/similarity-clustering>
- [5] Zhao, X., Liu, Y., & Sun, Y. (2021). An improved deep learning model for online ride-hailing service demand prediction. *Journal of Cleaner Production*, 315, 128202. <https://doi.org/10.1016/j.jclepro.2021.128202>
- [6] Data Preparation for Modelling <https://www.youtube.com/watch?v=Sf6jn8QZHhc>
- [7] <https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning>
- [8] <https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm>
- [9] <https://www.numpyninja.com/post/random-forest-classifier-a-beginner-s-guide>

PROOF READING WITH WRITING CENTRE

Writing Center Client Report




Berman, Aidan <aberm2@success.newhaven.edu>

To: ○ Gunda, Saiteja



Thu 5/4/2023 7:51 PM

[EXTERNAL SENDER]

 University of New Haven

Saiteja came into the Writing Center today to receive feedback on his paper for CSCI-6401. First, I read through his abstract and introduction. Next, I read through his proposed method and discussion, making minimal suggestions relating to grammar or punctuation. Next, I showed Saiteja how an APA reference section should be formatted, including the use of hanging indents, spacing, and alphabetical order. After discussing his paper for a few more minutes, he had no more questions and we wished each other a good rest of finals. Good luck, Saiteja!

[Learn more about Navigate for students, faculty and staff.](#)