Credit Card Fraud Detection with Machine Learning

Bekzod Mannapbekov
Student Number: 210872000
Supervisor Name: Eran Padumasada
MSc Big Data Science

Abstract— **It is crucial for banks to identify fraudulent credit card transactions and be updated since scams are evolving along with new technologies. The impact of fraud affects major areas such as people, banks, and the economy. Data science can be used to solve and prevent with the help of machine learning and well-structured datasets. This project attends to solve credit card fraud cases with machine learning algorithms through normalizing imbalanced data with different techniques. The challenge is to model past data where the fraud occurred and then apply the model to identify whether the new transactions are fraudulent or valid. The main objective is to get the highest accuracy possible while minimizing the wrong classification. This project applies four different machine learning models namely Naïve Bayes, Support Vector Machine, K-Nearest Neighbor, and Logistic Regression. The highest accuracy achieved in this project is with Logistic Regression 95.7%.**

Keywords— Machine Learning, Credit Card Fraud Detection, Data Cleaning, Data Exploration, Logistic Regression, Support Vector Machine, K-Nearest Neighbor, Naive Bayes.

I. INTRODUCTION

The definition of fraud in credit card transactions is the usage of an account without the consent of the legal card owner. In other words, credit card fraud (CCF) is when someone intentionally uses someone else's card for his/her personal reasons without the owner's or the issuing bank's notice (Oghenekaro and Ugwu, 2016). As pandemic restrictions are eased the number of CDF is rising significantly. In 2021 alone the amount stolen through fraudulent ways was over £1.3 billion and the highest percentage of it comes from authorized push payment fraud (APP) with £538.2 million in 2021. Where the victims are convinced by the scammers to make payments (the Guardian, 2022).

. In order to minimize and prevent these fraudulent practices, there are number of ways that can be researched and studied. The definition of fraud detection on the other hand is analysing the activities of users in order to identify, prevent or avoid fraudulent cases. In order to automate this issue, it demands the attention of modern technologies such as data science (Kou et al., 2004).

Assessing the problem from the data science point of view, in this type of classification problem the dataset is highly imbalanced since the number of fraud transactions is significantly less than the number of valid transactions (Makki et al., 2019). There are several ways of solving this issue starting from data mining techniques ending with artificial intelligence and one of the ways is through algorithms, which can learn from existing data patterns of valid as well as fraudulent transactions. When new transaction enters and if pattern is closer to fraudulent transactions then it shows *1* and if the pattern is closer to valid transactions, it returns *0*. **Fig. 1**. Shows the architecture diagram of the process (Seeja and Zareapoor, 2014).

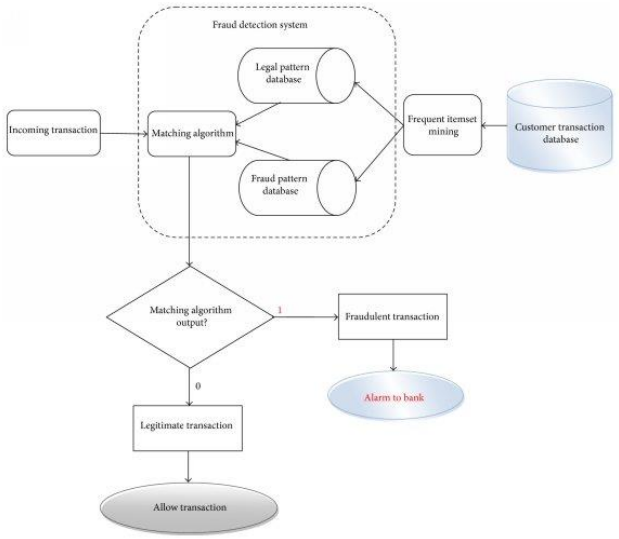

Fig. 1. The use of algorithms to identify patterns in transactions and classify into fraud or valid the transaction (Seeja and Zareapoor, 2014).

According to (Jolly, 2019), there are five common types of credit card fraud. Starting with Card-not-present fraud (CNP), this type of fraud occurs mainly from online transactions. Where the physical card is not needed and victims can just use the information on their cards and proceed payments. After, the unauthorized websites send an offer to people and when they accept the offer it steals the amount. Considering Australia, CNP fraud takes the lion share of credit card fraud cases with 85% and in 2018 the amount reached $477,920,701. Second type of CCF Counterfeit and skimming fraud, the amount victims lost in 2018 is $14,935,409 considering only Australia, but have significantly dropped over the years, when they announced new protection system made by chip technology. Skimming is a type of fraud where the fraudsters take over credit card information, and they sometimes use magnetic stripe or attach a device called credit card skimmer to ATMs. Another common type of CCF is lost and stolen card frauds, this type of fraud is self-explanatory, when the scammers get the full access of credit cards, they can use it until the card is frozen and the amount of money people lost in 2018 is $47,478,058. Card-never arrived fraud, when people order new card, it arrives through the mail most of the time and it can easily by pinched from the mail by fraudsters. The amount lost to this type of problem in 2018 is relatively smaller compared to other types and it is $6,231,308. The last common type is False application fraud, this occurs when fraudsters use

someone else's identity to get credit cards and the amount lost to this type of fraud in 2018 is $2,393,902.

II. BACKGROUND

A. *Importance of data preparation before machine learning process*

Data overall is the crucial aspect of machine learning processes and without data, it cannot operate, machine learning starts only after the clean dataset is created. There are problems with data being corrupted or it has a lot of missing data and duplicates and these issues result in inaccurate outcomes of machine learning predictions or classifications (Awan-Ur-Rahman, 2019). **Fig. 2.** illustrates simple data cleaning process, which is important in order to achieve accurate results.

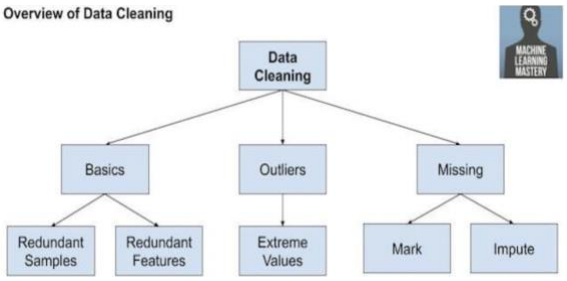

Fig. 2. Overview of Data Cleaning Process (Brownlee, 2020)

Outlier removal is a subjective practice but yet it is important to remove extreme outliers, due to it may violate analysis and assumptions. Handling missing data, there are a lot of ways of doing it and one of them is if the dataset is large enough and there are not many missing values it is easier to remove them. Handlining redundant data is vital since it negatively impacts predictive performance (Brownlee, 2020).

B. *The use of machine learning and types of algorithms*

To begin with machine learning algorithms, there are six main types of algorithms namely: Supervised learning, Unsupervised Learning, Semi-supervised learning, Reinforcement learning, Transduction and Learning to learn. In supervised learning algorithm, algorithm creates a function where inputs are mapped in a way that they give desired output. As an example of supervised learning is Classification problem, where the learner is required to learn from the examples and classify future outputs based on the previously learned data (Zhang, 2020). Unsupervised learning on the other hand, is a type of algorithm where it analyzes unlabeled data and cluster datasets. The way it works is, it finds hidden patters independently without any human intervention. As an example of unsupervised learning, there are many ways where this algorithm can be implemented, great fit for this type of algorithm is data-analysis, customer segmentation and image recognition (IBM Cloud Education, 2020). Semi-Supervised learning is a combination of both supervised and unsupervised learnings, it trains with labeled and unlabeled data but mostly unlabeled. Usually, it works with the dataset that has no outcome variables (English, 2019). Reinforcement learning is used when the machine learning models are trained to identify sequence. The agent learns, to accomplish goals in a very dynamic and complicated environment (Błażej Osiński, 2018). Transduction algorithm is similar to supervised learning; however, it does not create functions and focuses on predictions of new outputs by relying on new inputs, training inputs and training outputs. The last common type of machine learning algorithm is Learning to learn algorithm, where the main structure is built on own inductive bias based by relying on past experience (Zhang, 2020).

C. *Commonly used classification models in CCF Detection and their functions.*

According to (Bin Sulaiman, Schetinin and Sant, 2022), commonly used machine learning models particularly in CCF detection are Logistic Regression, Naïve bayes, K-Nearest Neighbor (KNN) and Support Vector Machine (SVM). They claim that KNN specifically in credit card fraud detection achieved 97.69% accuracy. Similarly, there were claims where KNN achieved only 72% in accuracy. Another research been conducted by (Alenzi and O, 2020) where Logistic regression accuracy reached 97.2% accuracy with 2.8% error rate. Another project, used Naïve Bayes model and achieved 94% accuracy (Husejinović, 2020). Considering research paper (Zhang, Bhandari and Black, 2020) SVM produced 99% accuracy in credit card fraud detection.

To begin with machine learning models, Logistic regression also known as the logit model is powerful in estimating the probability of an event with a dataset of independent variables. Dependent variables are bonded in binary meaning 0 and 1 since the end result is probability. Additionally, Linear regression is commonly used and main functionality compared to Logistic regression is that predicted Y can exceed the 0 and 1 range **Fig. 3**. (Brownlee, 2020). It is one of the most efficient ways to solve particular classification problem and the main advantages are it accurately predicts where the dataset is linearly separable and it is easy to implement and fast at solving the classification of unknown records (Grover, 2020). The downside of logistic regression is the problem with overfitting, in case the number of features is more than the number of observations it may result in overfitting (Raj, 2021).

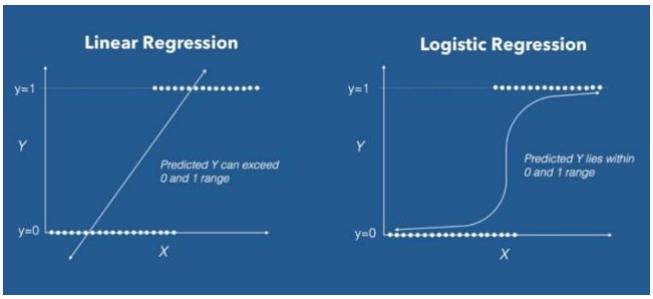

Fig. 3. Visual comparison of linear and logistic regression (Ayush Pant, 2019)

Naïve Bayes is useful with a large number of datasets and it is based on Bayes' Theorem with an assumption that one

specific feature in a class is not connected with any other feature. The formula of this model is as follows:

$$P(c|X) = \frac{P(x|c)P(c)}{P(x)} \quad (1)$$

In this equation (1) P(c|X) is posterior probability and class c is the target given predictor x attribute, followed by P(x|c) is a probability of predictor given class "likelihood", P(c) is the prior probability class and P(X) is the prior probability of predictor (Analytics Vidhya, 2017). The problem with this model is that all predictors are assumed to be independent and most of the time it is not the case (upGrad blog, 2021). **Fig. 4.** shows how it operates.

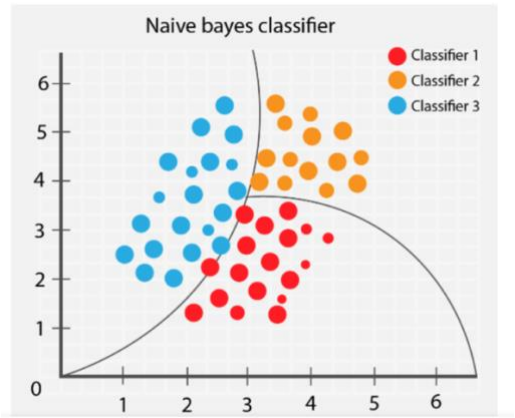

Fig. 4. Illustration of the model Navie Bayes (Analytics Vidhya, 2022)

K- Nearest Neighbor is a supervised machine learning algorithm and is popular in both regression and classification problems. The number of nearest neighbors to an unknown variable that needs to be predicted or classified depending on the problem then it is denoted by the symbol K. In other words, if an unknown variable is closer to Z it is assigned as Z the same structure with other variables. This algorithm mainly focused on distance metrics and most popular one is the *Minkowski* distance (Analytics Vidhya, 2021). One of the advantages of the algorithm is that tunning hyperparameters is simple since it requires only one value *K*. The downside is that with a large dataset and with greater dimensions it struggles (Soni, 2020). **Fig. 5.** Is the architecture diagram of the model.

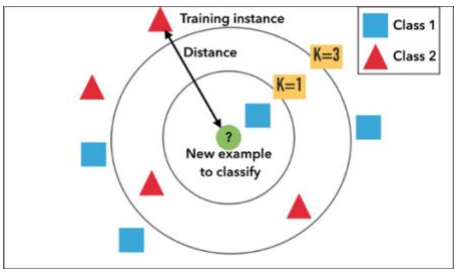

Fig. 5. K-Nearest Neighbor visual architecture diagram (Soni, 2020).

Support Vector Machine (SVM) is commonly used for classification problems, nevertheless, it can be applied to solve regression problems and it is a supervised machine learning algorithm. In this model, each data item is plotted with n-dimensional space (n is the number of features) and each coordinate has a particular value in a form of a feature. The next step is to start classification by identifying the hyper-plane that distinguishes classes. **Fig. 6** is the visual illustration of SVM.

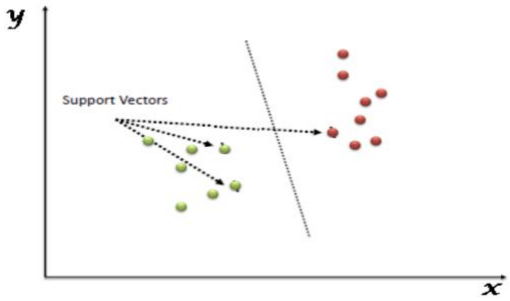

Fig. 6. Support Vector Machine visual interpretation (Analytics Vidhya, 2017).

SVM is particularly efficient when the margin of separation of classes is well contrast and the model is not suitable with large datasets (Analytics Vidhya, 2017).

D. Hyperparameter Tuning

According to (Weerts, Müller and Vanschoren, 2020), hyperparameter tuning often has more value than choosing a model. There are studies where the authors claim that leaving hyperparameters at their default values is the right approach, nonetheless, there are pieces of evidence that show improvements in results after tuning. Different machine learning models have different importance of hyperparameters, meaning every model has its own hyperparameter that needs to be tuned according to the importance rate.

While hyperparameters can be tuned manually, there are ways to automate the process with a help of the Optuna hyperparameter optimizer. The optimizer can quickly find the best hyperparameters for using machine learning models. It is easy to use meaning it does not interrupt codes and with few lines of code can be processed. There are additional features that come along with optimization and it includes python search space, quick visualization, and easy parallelization (optuna.readthedocs.io, n.d.).

E. Studies of various versions of fraud detection projects

Wide range of literatures used on this paper regarding fraud detection and are publicly available. Interesting literature (Bhardwaj and Gupta, 2016) provides information regarding techniques used in fraud detection and they are data mining applications and automated fraud detection. The use of data mining helps and unlocks raw data by turning it into useful information.

Another journal shows an example of telecommunications fraud detection approach with supervised and unsupervised learning. The project uses two types of clustering namely multilayer perceptron and the hierarchical agglomerative and in terms of neural networks analysis it performed well with only 2% in false positive and over 80% on true positive. Additionally, as part of their experimental results, they used ROC curve visualization. It is a graphical representation that displays separation of the number of overlapping distributions through identifying tradeoff between false positive and true positive rates by focusing on cut off points. Particularly, in ROC curve analysis in the visual diagram the area under the curve provides information about how the classier is preforming. Coming from area 0.5, it shows cases where it randomly classified and area 1 illustrates ideal classification. **Fig.7**. is the experimental results from the project and shows ROC curve diagram. (Hilas and Mastorocostas, 2008).

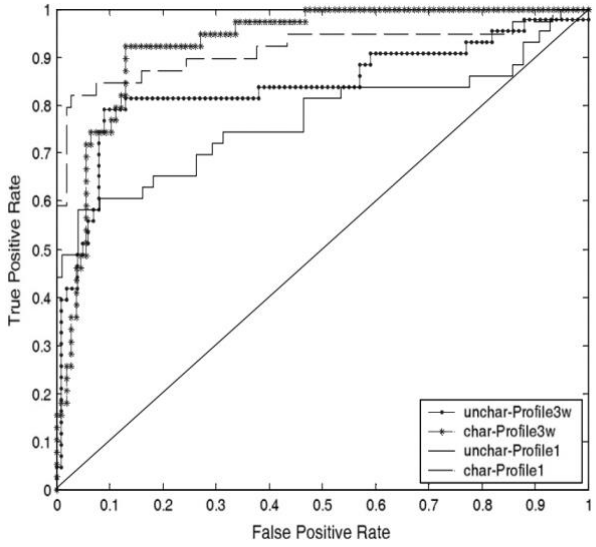

Fig. 7. ROC curve of supervised and unsupervised learning in credit card fraud detection (Hilas and Mastorocostas, 2008)..

Interesting research has been conducted similarly to the previous journal and it is about fraud detection in real-time messaging on social networks. They gathered data from various social media platforms as well as real case fraud events from Taiwan's ani-fraud center. With the real-world dataset they applied Natural Language Processing (NLP) to identify three major aspects "Chinese word segmentation, clearing all the stop words and separation symbols. Ultimately, they achieved fascinating accuracy with 80% accuracy rate (Chen et al., 2017).

There are some unique approaches to this issue and one of them is through complex network classification algorithm and it showed successful results in identifying illegal transactions. It was used on actual normal sized data set and by utilizing network reconstruction algorithm, which allows to create one instance from a reference group through creating representation of the deviation (Zanin et al., 2018).

There are studies where they took the problem from the different perspective by improving the alert feedback interaction. When the fraud is detected the system alerts and creates feedback in order to stop the transaction. **Fig. 8**. shows the data flow diagram in which there are three data mining engines such as: fraud detection database, customer/bank, credit card transaction database. Customer/bank database includes several options such as: withdrawal as well as deposit, opening of account operation transactions and credit card transactions. The actual database of fraud techniques provides detailed information about fraudulent actions towards costumer's credit card. Customers credit card history is stored inside database that contains all the previous transactions made by the customers (Asogwa, C and Chukwuneke, 2018).

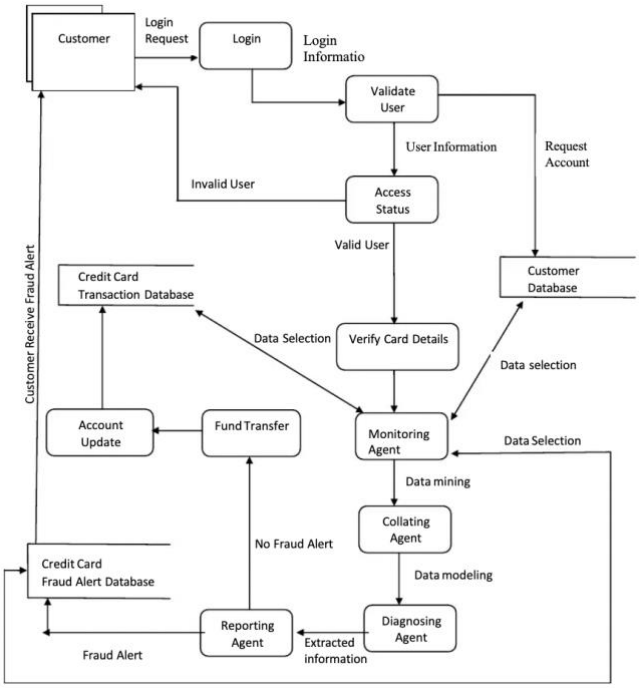

Fig. 8. Data flow diagram of the Alert system (Asogwa, C and Chukwuneke, 2018).

Blockchain is slowly but surely capturing financial areas and it has shown effectiveness in fraud detection. Blockchain reduces the complexity of the fraud detection process since in real time people share recorded data and transactions are only approved when all parties with the access to the recorded data approve it (Balagolla et al., 2021).

New technology called Hashgraph was found in mid 2010s by American computer scientist Leemon Baird. Hashgraph is considerably faster compared to blockchain technologies and might replace it in the future. It can process over thousands of transactions per second and most importantly securely (Schueffel, 2017).

III Methodology

To solve this problem the project uses the most effective machine learning algorithms obtained from the research papers. To begin with, the overall structure of this project is **Fig. 9.**

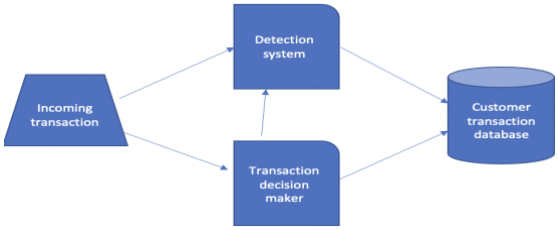

Fig. 9. The structure of credit card fraud detection

Starting with the overall methodology of this project, first thing first, it is important to understand the shape of the dataset before making further steps. Once the shape of the dataset is obtained, next step is data cleaning and exploration, which includes Sub Sampling in order to reduce computational cost as well as improve training process. Next step is feature engineering/data preprocessing and it includes correlations with visualizations, distribution of features and outliers. Then models (Logistic regression, SVM, KNN, Naïve Bayes) are executed. After accuracy of the models are obtained with their default hyperparameters, the next step is hyperparameter tuning with Optuna. The last step is model verification and assembling with cross validation, learning curve and AUC-ROC. In order to implement these steps, this project uses Jupyter Notebook with python 3.9 including variety of libraries that come with python.

A. Understanding the shape of the dataset

With regards to the dataset used in this project, the dataset was obtained from a website called Kaggle, the website is focused on data analysis and provides a wide range of datasets. The dataset follows ethical rules and no sensitive data is used in this project. The dataset consists of 31 columns and 28 of them are from v1 to v28 and the remaining are Time, Amount, and Class. Time illustrates the time between transactions and Amount is the amount of money transacted. The class consists of either 1 and 0, 1 stands for a fraudulent transaction, and 0 is a valid transaction. Data type of all the columns is Float 64 and range index is 284807 entries. **Fig. 10.** Displays the dataset using Python 3.9.

	Time	V1	V2	V3	V27	V28	Amount	Class
0	0.0	-1.359807	-0.072781	2.536347	0.133558	-0.021053	149.62	0
1	0.0	1.191857	0.266151	0.166480	-0.008983	0.014724	2.69	0
2	1.0	-1.358354	-1.340163	1.773209	-0.055353	-0.059752	378.66	0
3	1.0	-0.966272	-0.185226	1.792993	0.062723	0.061458	123.50	0
4	2.0	-1.158233	0.877737	1.548718	0.219422	0.215153	69.99	0
...	...	...	...	...	...	...	...	...
284802	172786.0	-11.881118	10.071785	-9.834783	0.943651	0.823731	0.77	0
284803	172787.0	-0.732789	-0.055080	2.035030	0.068472	-0.053527	24.79	0
284804	172788.0	1.919565	-0.301254	-3.249640	0.004455	-0.026561	67.88	0
284805	172788.0	-0.240440	0.530483	0.702510	0.108821	0.104533	10.00	0
284806	172792.0	-0.533413	-0.189733	0.703337	-0.002415	0.013649	217.00	0

Fig. 10. Credit Card Transaction Dataset

Considering mean values of the dataset, the Class and the Amount are strongly skewed. It is important to check for missing values since they can negatively impact machine learning models. There are multiple solutions but considering the size of this dataset it is easier to just remove them. However, there are no missing values in this dataset and the **Fig. 11**. displays it.

```
df0.columns[df0.isna().any()]
```

```
Index([], dtype='object')
```

Fig. 11. Output of missing data using python.

B. Cleaning data and exploration.

Next step is to understand the shape of the dataset and the **Fig. 12.** shows how imbalanced the data is since the number of valid transactions significantly higher than the number of fraudulent transactions. 0 is valid transactions with 284315.0 and 1 is fraudulent transaction with 492.0. Even though 492.0 is not a small number but compared to valid transactions number is barely visible on the count plot.

```
<AxesSubplot:xlabel='Class', ylabel='count'>
```

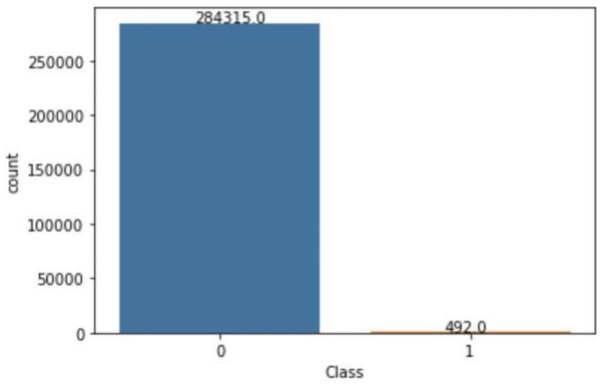

Fig. 12. Count plot of the number of valid transactions vs the number of fraudulent transactions.

Fig. 13. illustrates transactions within specific timeframe in order to understand what time has the most and the least number of transactions. According to the graph the greatest number of transactions are made during the day and the least during the night time.

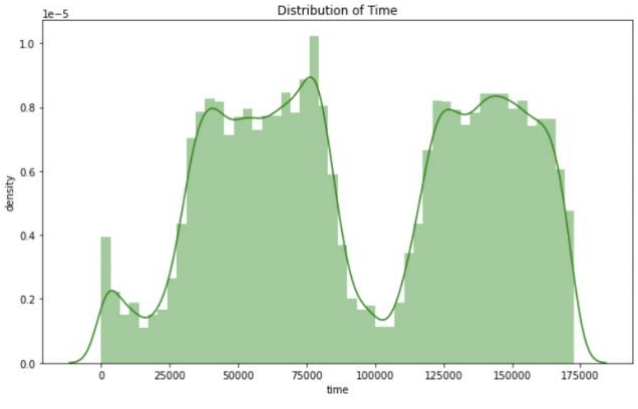

Fig. 13. Transaction within specific timeframe

Similarly, **Fig. 14.** shows the amount that was transacted and it shows that the lion share of the diagram takes rather smaller amounts and only few maximum amounts.

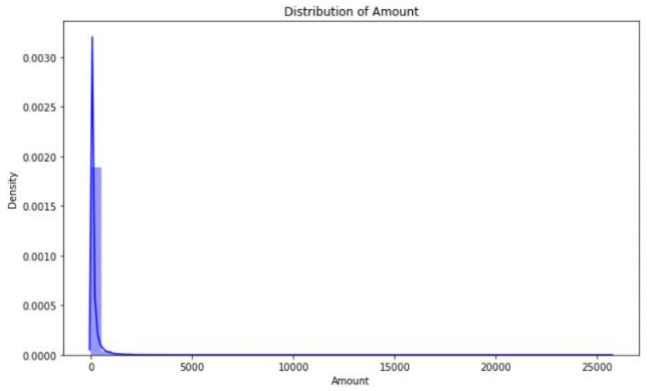

Fig. 14. Graph that shows Amount transacted each time.

Since the results gathered from the diagrams are extremely skewed, it is important to make subsampling. Subsampling is crucial to accurately evaluate as well as reduce computational cost (Krishnan and Srinivasan, 2022).

C. Sub-Sampling .

The objective in this case is to reach 1 to 1 with regards to fraudulent and valid transactions. To do that, all the valid transactions were reduced to 492 in randomized order since the number of fraud transactions are 492. **Fig. 15.** Shows the number of fraudulent transactions compared to valid transactions after sub-sampling.

```
<AxesSubplot:xlabel='Class', ylabel='count'>
```

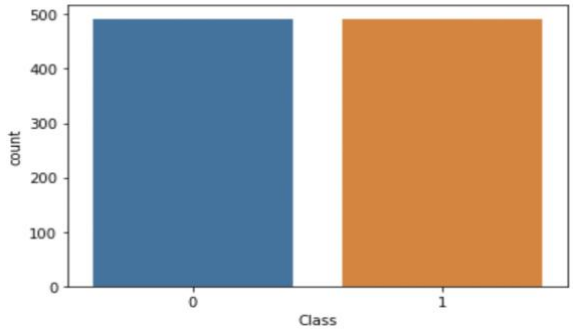

Fig. 15. Number of valid transactions compared to fraudulent transactions after sub-sampling

It is important to check whether the data is still skewed or not**.** New diagrams are used to plot distributions of scaled time and distributions of scaled amount. The information gathered shows that they are not as skewed before creating new scaled features.

D. Correlation Sub Sample Dataset

The next objective is to build visualized correlation from the subsampled dataset between class variables and predicting variables. From the heatmap **Fig. 16.** some raw data can be interpreted into information, there are two correlations positive and negative. Negative correlation is that lower the number higher the chances of the transaction being fraud (v2, v4, v11, v19) and positive correlation shows that higher the number higher the chances of the transaction being fraud (v10, v12, v14, v17).

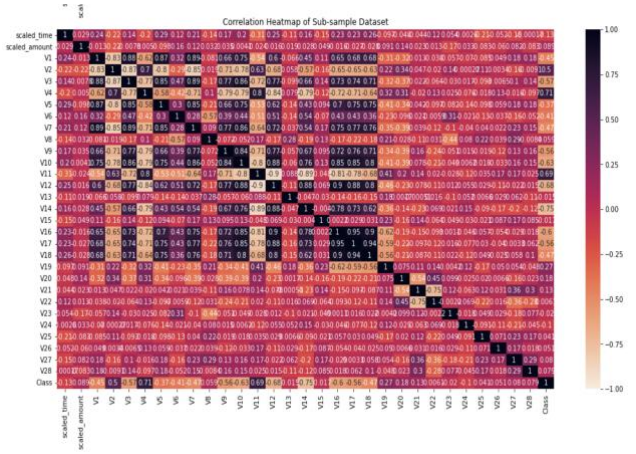

Fig. 16.. Correlation heatmap

The next step is distribution features, it is important to visually understand features with positive correlation and negative correlation. With the information gather from the heatmap, it is possible to plot V columns separately, V10 and V12 represent distribution features with negative correlation and V2 and V4 represent distribution of positive correlation **Fig. 17**.

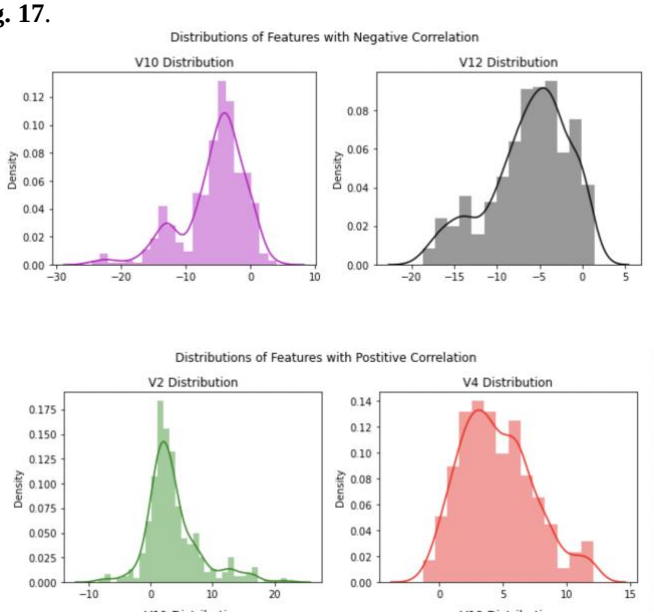

Fig. 17. Distribution features with positive and negative correlation

E. Analysing Outliers and removing extreme Outliers

Leading it to a field of outliers, there are abnormal features in datasets, while being an important aspect of statistical analysis it can be a problem. Outliers can interrupt analyses and violate outcomes (Jim, 2019). Since outliers directly impact the outcome of the machine learning models, some of the extreme outliers were carefully examined and removed from both positive and negative correlations to have a smoother prediction process. Then outliers were vissualised in a form of boxplot and some outliers are removed both from V2 and V10 since they had the highest number of outliers.

F. Model Building.

After understanding the shape of the data, making sub-sampling and cleaning unwanted data, the data is ready to run machine learning algorithms and classify fraudulent transactions. In order to run and test machine learning algorithms they were imported using sklearn. The next step is executing, Support Vector Machine, K-Nearest Neighbor, Logistic Regression and Naïve Bayes.

G. Hyperparameter Tuning

Once the accuracy score of each model is obtained, the next step is hyperparameter tunning with auto optimizer Optuna. Number of trails is set to 300 for each and every model except Naïve Bayes model due to it preformed the best with default hyperparameters, TABLE I illustrates the best hyperparameters for the models. Additional feature of Optuna is easy visualization tool that helps to quickly plot diagrams, in this case in order to obtain information about importance of hyperparameters of the models the importance parameter is plotted. With regards to KNN as discussed previously the only important hyperparameter in this model is n_neighbours meaning number of neighbours showing 99% and 1% is the algorithm. Logistic regression on the other hand, has two important parameters in order to perform well penalty has 57% importance rate and C has 43% importance rate. SVM the importance of Kernel is 63%, Gamma has 21% and Degree has 14%.

TABLE I

THE BEST HYPERPARAMETERS

Models	**Hyperparameters**
SVM	Karnel = Poly C = 0.13 Gamma = auto Degree = 3
Logistic Regression	Penalty = 12 C = 0.11 Solver = Liblinear
KNN	N_neaghbors = 2 Algorithm = kd_tree

IV. RESULTS

To begin with results, all the listed models are assessed in terms of total accuracy, cross validation, Learning Curve and ROC-AUC Curve as discussed in *Methodology* section.

A. Accuracy of the models.

The TABLE II shows that all the models preformed over 90% in accuracy score. However, the highest accuracy showed Logistic regression with 95.7% and the least accurate among the models is Naïve Bayes model with 91.94%. The second highest accuracy is KNN with 94.09% followed by SVM with 93.01%.

TABLE II

ACCURACY RESULTS OF THE MODELS

Models	**Accuracy**
SVM	93.01%
Naïve Bayes	91.94%
KNN	94.09%
Logistic Regression	95.7%

B. Accuracy results after hyperparameter tunning

After tunning hyperparameters, there small changes in accuracy TABLE II and starting with the highest accuracy, Logistic regression still preformed better with 95.8% followed by KNN with 95.1%. SVM showes 93.3% in accuracy after tunning. Naïve Bayes still remains at 91.94% since it prefers well with default parameters.

TABLE II

ACCURACY RESULTS AFTER HYPERPARAMETER TUNNING

Models	**Accuracy**
SVM	93.3%
KNN	95.1%
Logistic Regression	95.7%
Naïve Bayes	91.94%

C. Cross validation score of the models.

The TABLE IV shows cross validation score of the models, they are slightly lower than accuracy score and still the highest validation score showed Logistic regression with 92.92% and the lowest among the models is Naïve Bayes with 90.19%. SVM and KNN, however, shows the same validation accuracy 92.1%.

TABLE III

CROSS VALIDATION SCORE

Models	**Cross Validation**
SVM	92.1%
KNN	92.1%
Logistic Regression	92.92%
Naïve Bayes	90.19%

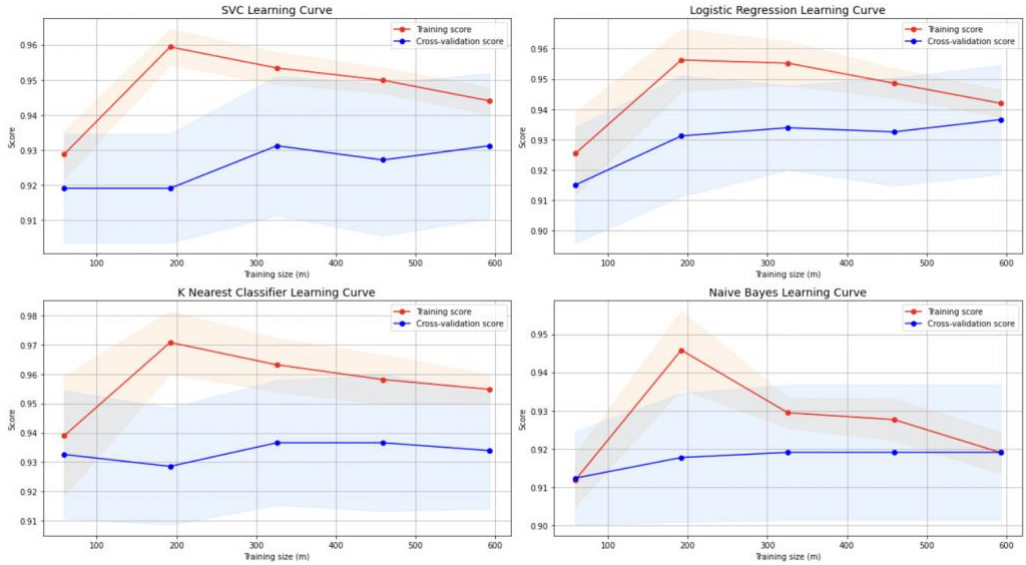

Fig. 18. Learning curve with training and cross validation of Support Vector Machine, Logistic Regression, K- Nearest Neighbor and Naïve Bayes Learning Curve.

D. Learning Curve Analysis..

Fig. 18. Illustrates learning curve of SVM, Logistic Regression, KNN and Naïve Bayes in terms of cross validation (blue line) and training score (red line). Among the models Logistic Regression preformed the best in terms of learning curve, followed by SVM and KNN they performed relatively the same. Least accurate in learning curve is Naïve Bayes

D. ROC-AUC Score

TABLE IV shows the ROC-AUC score of the models, starting with the highest Logistic regression with 97.4% and the least accurate is Naïve Bayes with 90.7%. Second highest is SVM wirh 96.6% followed by KNN with 91.7%.

TABLE IV

ROC-AUC SCORE

Models	**ROC-AUC**
SVM	96.6%
KNN	91.7%
Logistic Regression	97.4%
Naïve Bayes	90.7%

V. CONCLUSION AND DISCUSSIONS

Considering the fact that fraudulent cases in credit card is increasing every year, this project developed four most efficient machine learning models to solve the issue with fraudulent transactions. All the models preformed above 90% in all testing methods listed above and Logistic regression in this problem preformed the best compared to other models. Meanwhile, SVM and KNN preformed relative the same in every testing method. The least accurate model is Naïve Bayes.

The way to improve this project in the future is to get larger dataset and train models and learn new behaviors of fraudulent transactions. At this moment this project shows high accuracy but if it is executed in real-world bank with current dataset, it would not perform well since the dataset used in this project is not suitable in actual banks with more data it can improve and adapt so it can stop more fraudulent transactions. It all comes to data and collaboration with banks would significantly improve this project by having an access to wide number of transactions.

V. Acknowledgement

First of all, I would like to thank my supervisor senior Lecturer Eran Padumasada for the support he provided throughout the process. I would like to thank Dr.Sangeetha Elango for providing information about finance in banks. I would like to show my deepest gratitude to my family for supporting me throughout my university time and ultimately, my classmates for sharing their thoughts and ideas.

REFERENCE

Akers, M. and Gissel, J. (2006). What Is Fraud and Who Is Responsible? Journal of Forensic Accounting, [online] 7(1), pp.247–256. Available at: https://core.ac.uk/download/pdf/213082859.pdf

the Guardian. (2022). UK victims lost £1.3bn in 2021 amid surge in online fraud, new data shows. [online] Available at: https://www.theguardian.com/money/2022/jun/29/uk-victims-lost-13bn-in-2021-amid-surge-in-online-new-data-shows.

Kou, Y., Lu, C.-T., Sirwongwattana, S. and Huang, Y.-P. (2004). Survey of fraud detection techniques. [online] IEEE Xplore. Available at: doi:10.1109/ICNSC.2004.1297040.

Debray, T. (n.d.). Classification in Imbalanced Datasets. [online] Available at:https://www.academia.edu/54903003/Classification_in_Imbalance_Datasets

Jolly, W. (2019). 5 most common types of credit card frauds explained. [online] Savings.com.au. Available at: https://www.savings.com.au/credit-cards/credit-card-fraud.

Seeja, K.R. and Zareapoor, M. (2014). FraudMiner: A Novel Credit Card Fraud Detection Model Based on Frequent Itemset Mining. The Scientific World Journal, [online] 2014, pp.1–10. doi:10.1155/2014/252797

Bhat K, S.K. (2020). Credit Card Fraud Detection using Machine Learning Methods. International Journal for Research in Applied Science and Engineering Technology, 8(6), pp.1436–1440. doi:10.22214/ijraset.2020.6233.

Awan-Ur-Rahman (2019). What is Data Cleaning? How to Process Data for Analytics and Machine Learning Modeling? [online] Medium. Available at: https://towardsdatascience.com/what-is-data-cleaning-how-to-process-data-for-analytics-and-machine-learning-modeling-c2afcf4fbf45 [Accessed 5 July.. 2022].

Brownlee, J. (2020). Tour of Data Preparation Techniques for Machine Learning. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/data-preparation-techniques-for-machine-learning/ [Accessed 6 July. 2022].

Hpe.com. (2022). What is Machine Learning? | Glossary. [online] Available at:https://www.hpe.com/uk/en/what-is/machine-learning.html?jumpid=ps_b64ed7xwyq_aid-520061736&ef_id=Cj0KCQjwxb2XBhDBARIsAOjDZ34cSflQRkxXStEEU1pGkeWtd9sgVUN52pDpSLfquosqSMglGM5MTooaAszCEALw_wcB:G:s&s_kwcid=AL [Accessed 12 June. 2022].

Chakraborty, Chiranjit and Joseph, Andreas, Machine Learning at Central Banks (September 1, 2017). Bank of England Working Paper No. 674, Available at SSRN: https://ssrn.com/abstract=3031796 or http://dx.doi.org/10.2139/ssrn.3031796

Zhang, Y. (2010). New Advances in Machine Learning. [online] Google Books. BoD – Books on Demand. Available at: https://books.google.co.uk/books?hl=en&lr=&id=XAqhDwAAQBAJ&oi=fnd&pg=PA19&dq=types+of+machine+learning+problems&ots=r2LiaUzeKt&sig=zr5ZnWV8G9rc1f9DxuEyqG_tl-o#v=onepage&q=types%20of%20machine%20learning%20problems&f=false [Accessed 13 Aug. 2022].

IBM Cloud Education (2020). What is Unsupervised Learning? [online] www.ibm.com. Available at: https://www.ibm.com/cloud/learn/unsupervised-learning.

English. (2019). DataRobot Automated Machine Learning. [online] Available at: https://www.datarobot.com/wiki/semi-supervised-machine-learning/.

Błażej Osiński (2018). What is reinforcement learning? The complete guide - deepsense.ai. [online] deepsense.ai. Available at: https://deepsense.ai/what-is-reinforcement-learning-the-complete-guide/.

Bin Sulaiman, R., Schetinin, V. and Sant, P. (2022). Review of Machine Learning Approach on Credit Card Fraud Detection. Human-Centric Intelligent Systems. doi:10.1007/s44230-022-00004-0.

Alenzi, H.Z. and O, N. (2020). Fraud Detection in Credit Cards using Logistic Regression. International Journal of Advanced Computer Science and Applications, [online] 11(12). doi:10.14569/ijacsa.2020.0111265.

Husejinović, A. (2020). Credit card fraud detection using naive Bayesian and C4.5 decision tree classifiers. [online] 8(1), pp.1–5. Available at: https://core.ac.uk/download/pdf/287201129.pdf [Accessed 13 Aug. 2022].

Jolly, W. (2019). 5 most common types of credit card frauds explained. [online] Savings.com.au. Available at: https://www.savings.com.au/credit-cards/credit-card-fraud.

Zhang, D., Bhandari, B. and Black, D. (2020). Credit Card Fraud Detection Using Weighted Support Vector Machine. Applied Mathematics, 11(12), pp.1275–1291. doi:10.4236/am.2020.1112087.

English. (2019). DataRobot Automated Machine Learning. [online] Available at: https://www.datarobot.com/wiki/semi-supervised-machine-learning/.

Błażej Osiński (2018). What is reinforcement learning? The complete guide - deepsense.ai. [online] deepsense.ai. Available at: https://deepsense.ai/what-is-reinforcement-learning-the-complete-guide/.

Bin Sulaiman, R., Schetinin, V. and Sant, P. (2022). Review of Machine Learning Approach on Credit Card Fraud Detection. Human-Centric Intelligent Systems. doi:10.1007/s44230-022-00004-0.

Alenzi, H.Z. and O, N. (2020). Fraud Detection in Credit Cards using Logistic Regression. International Journal of Advanced Computer Science and Applications, [online] 11(12). doi:10.14569/ijacsa.2020.0111265

Husejinović, A. (2020). Credit card fraud detection using naive Bayesian and C4.5 decision tree classifiers. [online] 8(1), pp.1–5. Available at: https://core.ac.uk/download/pdf/287201129.pdf [Accessed 13 Aug. 2022].

Zhang, D., Bhandari, B. and Black, D. (2020). Credit Card Fraud Detection Using Weighted Support Vector Machine. Applied Mathematics, 11(12), pp.1275–1291. doi:10.4236/am.2020.1112087

Grover, K. (2020). Advantages and Disadvantages of Logistic Regression. [online] OpenGenus IQ: Learn Computer Science. Available at: https://iq.opengenus.org/advantages-and-disadvantages-of-logistic-regression/.

Ayush Pant (2019). Introduction to Logistic Regression. [online] Medium. Available at: https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148.

Analytics Vidhya. (2022). Building Naive Bayes Classifier from Scratch to Perform Sentiment Analysis. [online] Available at: https://www.analyticsvidhya.com/blog/2022/03/building-naive-bayes-classifier-from-scratch-to-perform-sentiment-analysis/ [Accessed 13 Aug. 2022].

Brownlee, J. (2020). 4 Types of Classification Tasks in Machine Learning. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/types-of-classification-in-machine-learning/.

Wolff, R. (2020). Classification Algorithms in Machine Learning: How They Work. [online] MonkeyLearn Blog. Available at: https://monkeylearn.com/blog/classification-algorithms/

www.ibm.com. (n.d.). What is Logistic regression? | IBM. [online] Available at: https://www.ibm.com/uk-en/topics/logistic-regression#:~:text=Logistic%20regression%20estimates%20the%20probability [Accessed 23 Jun. 2022].

Raj, A. (2021). The Perfect Recipe for Classification Using Logistic Regression. [online] Medium. Available at: https://towardsdatascience.com/the-perfect-recipe-for-classification-using-logistic-regression-f8648e267592#:~:text=Logistic%20regression%20is%20easier%20to

Rout, A.R. (2020). Advantages and Disadvantages of Logistic Regression. [online] GeeksforGeeks. Available at: https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/.

Analytics Vidhya. (2017). Learn Naive Bayes Algorithm | Naive Bayes Classifier Examples. [online] Available at: https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/#:~:text=Naive%20Bayes%20Model- [Accessed 21 Feb. 2021].

upGrad blog. (2020). Naive Bayes Explained: Function, Advantages & Disadvantages, Applications in 2020. [online] Available at: https://www.upgrad.com/blog/naive-bayes-explained/.

Analytics Vidhya. (2021). KNN - The Distance Based Machine Learning Algorithm. [online] Available at: https://www.analyticsvidhya.com/blog/2021/05/knn-the-distance-based-machine-learning-algorithm/#:~:text=The%20abbreviation%20KNN%20stands%20for

www.cs.cornell.edu. (n.d.). Lecture 2: k-nearest neighbors / Curse of Dimensionality. [online] Available at: https://www.cs.cornell.edu/courses/cs4780/2018fa/lectures/lecturenote02_kNN.html#:~:text=The%20k-nearest%20neighbor%20classifier

Analytics Vidhya. (2017). SVM | Support Vector Machine Algorithm in Machine Learning. [online] Available at: https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/#h2_1 [Accessed 8 Aug. 2022].

Weerts, H., Müller, A. and Vanschoren, J. (2020.). Importance of Tuning Hyperparameters of Machine Learning Algorithms. [online] Available at: https://arxiv.org/pdf/2007.07588.pdf [Accessed 10 July. 2022].

optuna.readthedocs.io. (n.d.). Optuna: A hyperparameter optimization framework — Optuna 2.10.1 documentation. [online] Available at: https://optuna.readthedocs.io/en/stable/ [Accessed 8 Aug. 2022].

Bhardwaj, A. and Gupta, R. (2016). Financial Frauds: Data Mining based Detection – A Comprehensive Survey. International Journal of Computer Applications, 156(10), pp.20–28. doi:10.5120/ijca2016912538.

Hilas, C.S. and Mastorocostas, P.As. (2008). An application of supervised and unsupervised learning approaches to telecommunications fraud detection. Knowledge-Based Systems, 21(7), pp.721–726. doi:10.1016/j.knosys.2008.03.026.

CHEN, L.-C., HSU, C.-L., LO, N.-W., YEH, K.-H. and LIN, P.-H. (2017). Fraud Analysis and Detection for Real-Time Messaging Communications on Social Networks. IEICE Transactions on Information and Systems, E100.D(10), pp.2267–2274. doi:10.1587/transinf.2016ini0003.

Zanin, M., Romance, M., Moral, S. and Criado, R. (2018). Credit Card Fraud Detection through Parenclitic Network Analysis. Complexity, 2018, pp.1–9. doi:10.1155/2018/5764370.

Asogwa, D.C., C, A.B. and Chukwuneke, C.I. (2018). Credit Card Fraud Detection System Using Intelligent Agents and Enhanced Security Features. www.academia.edu. [online] Available at: https://www.academia.edu/37103019/Credit_Card_Fraud_Detection_System_Using_Intelligent_Agents_and_Enhanced_Security_Features [Accessed 5 Aug. 2022].

Balagolla, E.M.S.W., Fernando, W.P.C., Rathnayake, R.M.N.S., Wijesekera, M.J.M.R.P., Senarathne, A.N. and Abeywardhana, K.Y. (2021). Credit Card Fraud Prevention Using Blockchain. [online] IEEE Xplore. doi:10.1109/I2CT51068.2021.9418192.

Schueffel, P. (2017). Alternative Distributed Ledger Technologies Blockchain vs. Tangle vs. Hashgraph - A High-Level Overview and Comparison -. SSRN Electronic Journal. [online] doi:10.2139/ssrn.3144241.

Krishnan, M. and Srinivasan, M.K. (2022). Credit Card Fraud Detection: An Exploration of Different Sampling Methods to Solve the Class Imbalance Problem. Algorithms for Intelligent Systems, pp.825–837. doi:10.1007/978-981-16-5747-4_71.

Mondal, I.A., Haque, Md.E., Hassan, A.-M. and Shatabda, S. (2021). Handling Imbalanced Data for Credit Card Fraud Detection. [online] IEEE Xplore. doi:10.1109/ICCIT54785.2021.9689866.

Hilbers, A.P., Brayshaw, D.J. and Gandy, A. (2019). Importance subsampling: improving power system planning under climate-based uncertainty. Applied Energy, 251, p.113114. doi:10.1016/j.apenergy.2019.04.110.

Frost, J. (2019). Guidelines for Removing and Handling Outliers in Data - Statistics By Jim. [online] Statistics By Jim. Available at: https://statisticsbyjim.com/basics/remove-outliers/.

Soni, A. (2020). Advantages And Disadvantages of KNN. [online] Medium. Available at: https://medium.com/@anuuz.soni/advantages-and-disadvantages-of-knn-ee06599b9336.

Makki, S., Assaghir, Z., Taher, Y., Haque, R., Hacid, M.-S. and Zeineddine, H. (2019). An Experimental Study With Imbalanced Classification Approaches for Credit Card Fraud Detection. IEEE Access, 7, pp.93010–93022. doi:10.1109/access.2019.2927266.

.

APENDIX

A. *Figures*

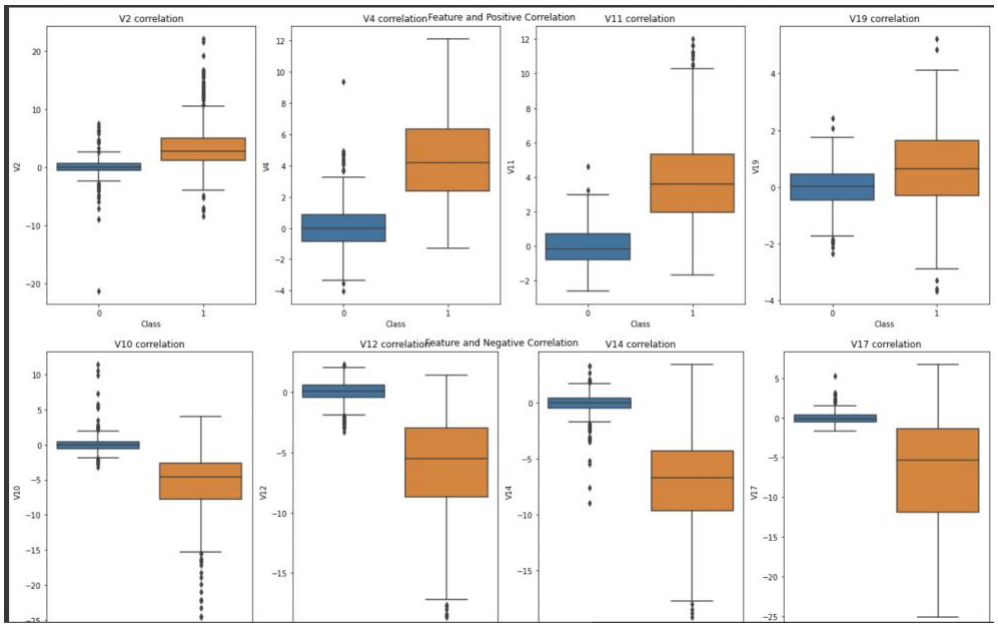

Fig. A. 1. Outliers of the dataset positive correlation and negative correlation

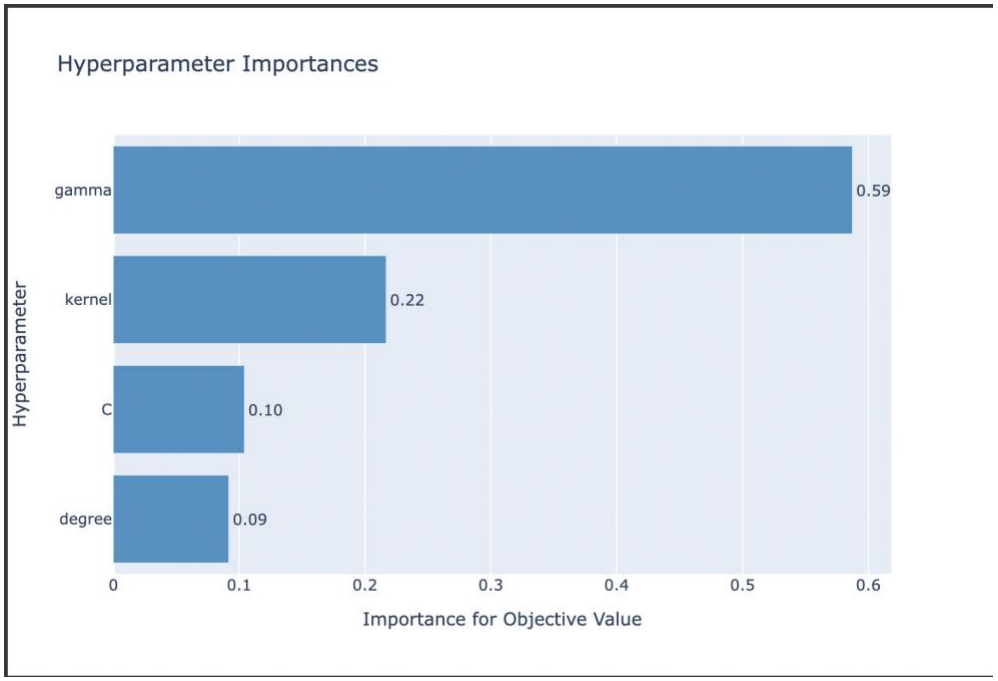

Fig. A. 2. SVM Importance of hyperparameters visualization with Optuna

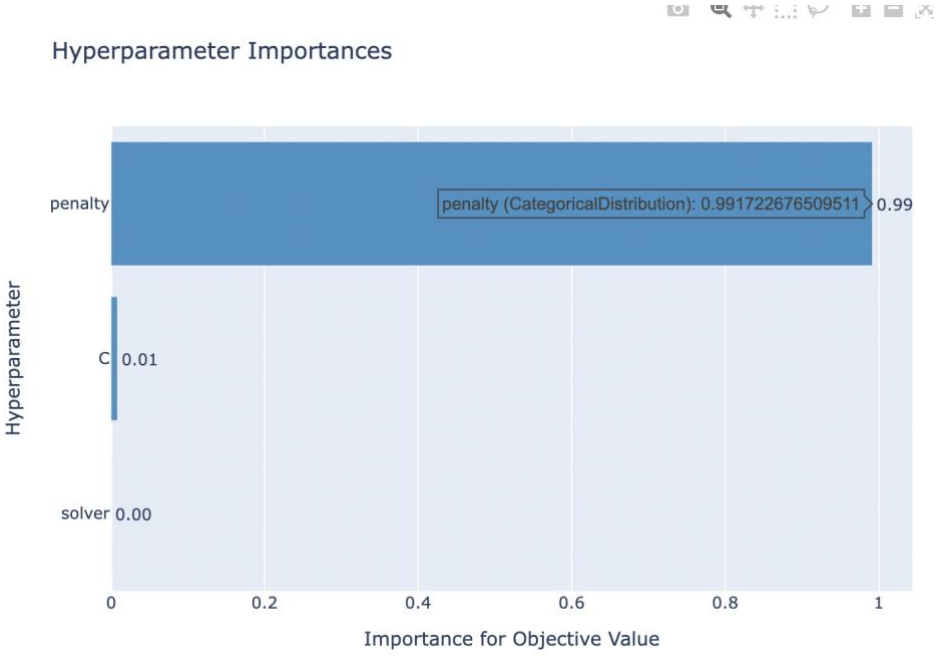

Fig. A. 3. Logistic Regression importance of hyperparameters visualization with Optuna

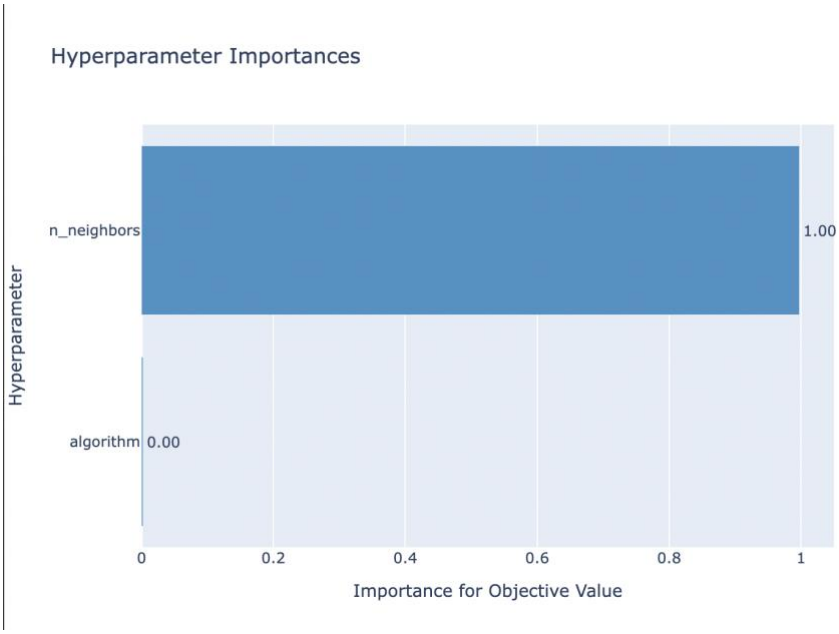

Fig. A. 4. KNN importance of hyperparameters visualization with Optuna

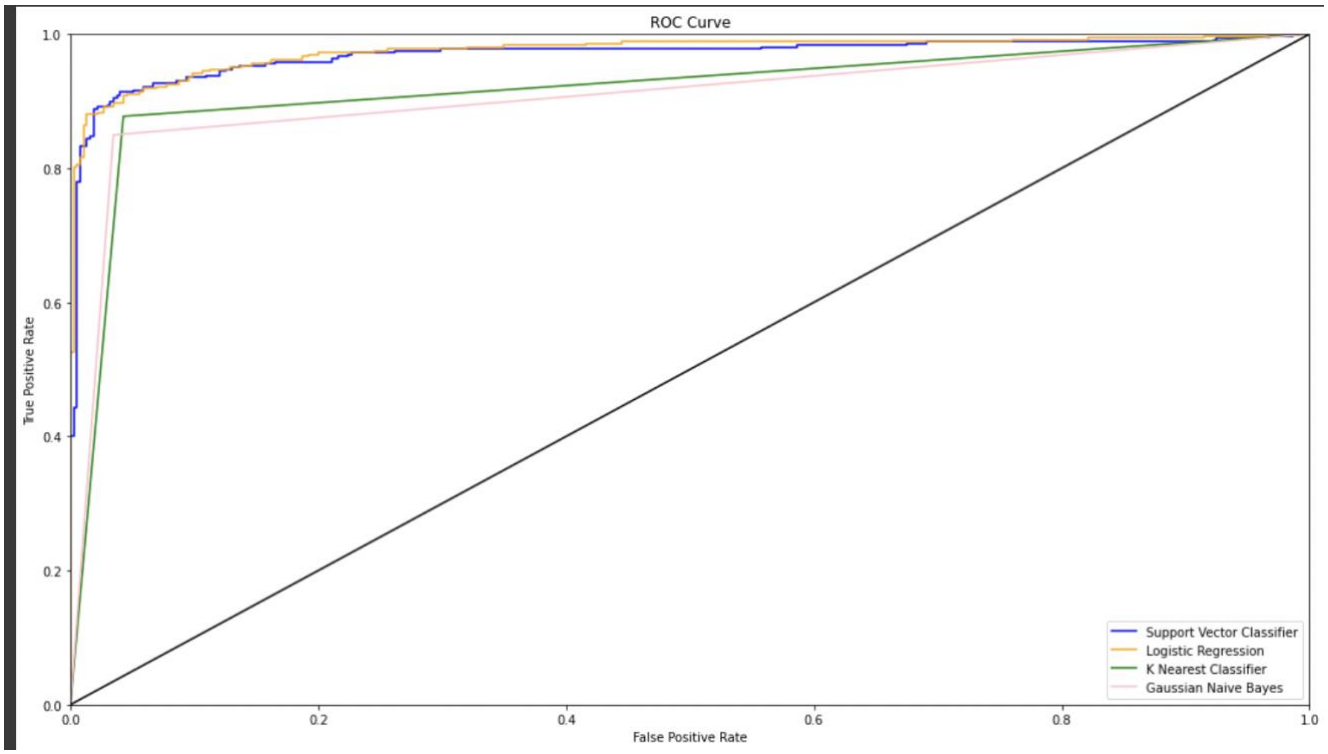

Fig. A. 4. ROC-AUC curve of Logistic Regression, KNN, Naïve Bayes, SVM.

MSc Project - Reflective Essay

Project Title:	Credit card Fraud Detection with Machine Learning
Student Name:	Bekzod Mannapbekov
Student Number:	210872000
Supervisor Name:	Eran Padamusada
Course of Taken:	MSc Big Data Science

Introduction.

This project addresses the problem regarding fraudulent transactions and provides solutions. The project illustrates machine learning and data mining techniques to solve the issue. Fraud in finance is a serious issue in the current generation, some forms of credit card fraud cases are not recorded and despite that the data gathered from the researches illustrates £1.3 billion every year, which is more likely less than the actual figure since not all the cases are recorded (the Guardian 2022).

After deciding with my project idea, the way I approached this project, I made research about machine learning in specific aspects such as: types of problems, what models suite specific type of problem and problems with the results. With the information gathered, I researched further about data in machine learning, importance, structure and issues with data. Then I reviewed previously made projects where other researchers attempted the problem. After reviewing the projects, I outlined all the highest accuracies achieved. Then I made a list of machine learning models with the highest accuracies according to previous works. Additionally, I made research about fraud detection in different fields. Then I constructed methodology, step by step approach to implement the idea and downloaded data from an open-source website Kaggle. Then I executed the models by using the knowledge gathered from the research papers and lecture notes from the two semesters and tested with different testing methods and the highest accuracy I achieved is 95.7% in classifying fraudulent transactions with Logistic Regression.

This reflective essay supplements research paper with additional information by elaborating points covered above. The additional information covers analysis of strength and weaknesses, several possible ways of improving current machine learning fraud detection approach, critically analyses the connection between practical work and theory work and concludes with ethical, social, legal and sustainability issues within the scope of this project.

Analyses of Strengths and Weaknesses.

Main strengths of this project, it uses the most effective machine learning algorithms gathered from several previously made projects and for the given dataset it shows very good accuracy (Logistic regression 95.7% with 4.3% wrong classification rate). It automates the process of detecting fraud transactions in result, it significantly reduces time in solving this issue, since manually approach would take a lot more time and effort. It opens up new creative and complex ways to solve the issue, not only in banks but in different fields such as ecommerce websites and blockchain. Considering the fact that most of the credit card fraud cases are registered under authorised push payment fraud (APP) where in 2021 it captured over £500 million, the approach used on this project counters it and potentially have an impact on reducing the APP cases.

In terms of weaknesses, this project as in this stage is not ready to be used in real-world banks, since the amount of data used to train the models is not near comparable with actual sized banks dataset. To effectively, use these machine learning algorithms, banks need a full team of data scientists to handle data flow and keep the fraud detection system up to date. Coming back to the types of financial frauds, from my personal understanding, considering this paper, if the credit card gets stolen, it is less likely to detect all the transactions made by the thieves before the legal owner of the card freezes the account. Additionally, if the scammers use someone else's information to open an account this particular approach cannot detect it and it demands a new approach with different security features in a form of prove of address, verification of personal information and more. It was quite challenging to clean the dataset, since it was incredible skewed and balance outlier removal stage, where I had to decide how much I need to remove.

Possibilities for further work

The project can be improved by adding more data, ideally with actual data from a bank if it follows ethical and legal rules. It would benefit the machine learning models if the new dataset has more elements such as: Location and IP address.

If given more time, I would personally focus my attention towards blockchain and integrate the implemented machine learning algorithms with it. I would attempt in creating distributed database, where it consists of a loop that consistently expands number of ordered records and the records are connected to a cryptography. Every record shares the cryptography of the previous record. It is interesting to know what would happen if it connects with my already implemented machine learning models. The reason of choosing blockchain is that considering Indian banks, they are currently facing challenges in increasing operational cost, the number of fraudulent transactions and insuring transparency. On the other hand, blockchain technologies are evolving in rapid phase and disturbing traditional systems. First of all, blockchain is the base of majority of cryptocurrencies and bitcoin one of them. The biggest banks in India are currently testing various blockchain applications in fields such as: vendor financing, syndicated loans and more (Garg et al., 2020).

Critical analysis of the relationship between practical work and theory.

To begin with, most of the aspects from the theory part of the project aligns with the practical outcome of the project. However, there are some points I would like to share that I did not expect. One of them is the impact of outliers (Brownlee, 2020), the effect of extreme outliers on accuracy was not strict in this case. However, when I executed the models with and without removing extreme outliers, the affect was relatively small only 0.4%. Maybe the reason behind it is the sub-sampling used in this project, with rather larger dataset the impact might be bigger.

Another aspect, regarding Naïve Bayes machine learning model (Analytics Vidhya, 2022), from the theory part I was expecting the model to perform the highest accuracy or at least second best, since the model performs well with large datasets, during the testing stage the model performed the lowest among other models but still over 90%.

The last aspect is hyperparameter tunning (Weerts, Müller and Vanschoren, 2020), after reading and researching, I was expecting tuning hyperparameters will lead to significant improvements in accuracy score but in result, the improvements were barely noticeable and Naïve Bayes model performance was reduced after tunning.

Ethical, social, legal and sustainability issues of the project.

To complete this project, I used Google Collab research platform, and Google as tech company is net zero meaning it is carbon neutral since 2017. The company is claiming to operate with carbon free by 2030 (BBC news, 2020). Regarding legal and ethical issues in this project, the dataset is obtained from open-source platform Kaggle and does not involve any personal information. The data used in this project

is fully anonymous. This research paper do not have any misleading information and the results do not cause any harm.

Personal improvements and benefits.

This project opened up new perspectives in science overall for me, I could never imagine just a year ago that I could accomplish a project like this. It gave me an ability to truly understand life cycle of data science starting from difference between Artificial intelligence and Machine Learning to understanding businesses where the main products are Artificial Intelligence and Machine Learning. Being able to debate in data science and build constructive methods to solve real-world problems. I would like to show my deepest gratitude to all the lectures who explained all the subjects I attended with great volume of high quality content.

Reference

Garg, P., Gupta, B., Chauhan, A.K., Sivarajah, U., Gupta, S. and Modgil, S. (2020). Measuring the perceived benefits of implementing blockchain technology in the banking sector. Technological Forecasting and Social Change, 163, p.120407. doi:10.1016/j.techfore.2020.120407.

BBC New. (2020). s Google says its carbon footprint is now zero. [online] 14 Sep. Available at: https://www.bbc.co.uk/news/technology-54141899.

the Guardian. (2022). UK victims lost £1.3bn in 2021 amid surge in online fraud, new data shows. [online] Available at: https://www.theguardian.com/money/2022/jun/29/uk-victims-lost-13bn-in-2021-amid-surge-in-online-new-data-shows.

Brownlee, J. (2020). Tour of Data Preparation Techniques for Machine Learning. [online] Machine Learning Mastery. Available at: https://machinelearningmastery.com/data-preparation-techniques-for-machine-learning/ [Accessed 25 July. 2022].

Analytics Vidhya. (2017). Learn Naive Bayes Algorithm | Naive Bayes Classifier Examples. [online] Available at :https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/#:~:text=Naive%20Bayes%20Model- [Accessed 28 July. 2021].

Weerts, H., Müller, A. and Vanschoren, J. (2020.). Importance of Tuning Hyperparameters of Machine Learning Algorithms. [online] Available at: https://arxiv.org/pdf/2007.07588.pdf [Accessed 10 July. 2022].