Capstone Project in Data Science

Fraud Detection in Electricity and Gas Consumption

PROJECT REPORT

COURSE INSTRUCTOR: DR SUMAN SAHA

COLLABORATORS
ZAKRIA SAAD
TINASHE HAFE
TRYMORE NCUBE

ABSTRACT

This report addresses the challenge of detecting fraud in electricity consumption in developing countries, where conventional meters are the norm due to their affordability. Time series data-logging is a useful tool in detecting irregularities in consumption patterns, as deviations from regular consumption can be easily identified. However, detecting fraud with conventional meters is difficult due to the hidden anomalies within normal consumption patterns. In this study, a combination of large datasets from consumers and invoice data in Tunisia are investigated using several Machine Learning (ML) classification algorithms. Extensive feature engineering, including multivariate Gaussian distribution, is performed to improve the efficiency of the ensemble classifiers, particularly the Light Gradient Boosting (LGB) algorithm. The LGB algorithm outperforms other algorithms and achieves realistic performance even with challenging, unbalanced, and uncorrelated input datasets.

INTRODUCTION

Fraudulent consumption is usually registered in countries with poor communities and increases the already high technical losses. Although the number of fraudulent consumers is not large, they create significant losses and encourage similar behavior. The Tunisian Company of Electricity and Gas (STEG https://www.steg.com.tn/en/institutionnel/mission.html), responsible for delivering electricity and gas in Tunisia, has encountered major losses of about 200 million Tunisian Dinars, from electricity fraud. The solution provided in this paper will enhance utility companies' revenues and reduce the losses caused by electricity fraud.

Over 135,000 of consumers and their consumption details totalling 4.5 million records were imported from the datasets offered by. On average, 33 invoices per consumer, from 1977 to 2019, were issued out of the normal monthly sequence. The aim of this work is to detect and recognize consumers with suspicious consumption, using consumer billing history, reading remarks, statue of the counter and intrinsic characteristics.

Daily or hourly consumption is easy to analyze, as the periodicity and variability of consumption are computed from the data time series. However, with monthly readings, it is impossible to catch these indicators as usually thieves consume at normal rate for half of the month, while for the other half, the counters are tampered with, thus on average the consumption is diminished, but without creating suspicion. The thieves have a weakness in that they don't collaborate and are unaware of each other's consumption behaviors. Therefore, there is no regular pattern and the variability is usually higher. In contrast, normal consumers have a regular pattern of electricity consumption, according to the day of the week and holidays.

In most developing countries, where fraudulent behavior is more frequent, conventional meters are still in place, thus, it is not possible to extract the daily and weekly periodicity of consumption. The variability of consumption is also lost, since the total consumption is read, on average, once a month.

The only reliable features, in case of conventional counters, are the reading remarks, counter status, type of tariff, meter, invoice date and monthly consumption level, that are very weakly correlated with the target or flag associated with each consumer. Another peculiarity of the input dataset is its unbalanced flag, as only a few consumers are thieves, whereas the majority are normal. From 135,493 consumers, 7,566 are targeted as thieves, representing 5.58% of the total number of consumers..

The datasets provided by the Tunisian utility company for training and testing have major issues from the classification point of view, due to the fact that many identical meter numbers are present in both datasets. Furthermore, there are counter numbers with different values of flag. Therefore, the classifier will provide misleading and ambiguous results as it recognizes, in the testing phase, the counter numbers from training, with the same or a different flag. Hence, after verifying the datasets, we decided to use only the training sets that are generous in volume and with unique identifiers, to avoid misleading results.

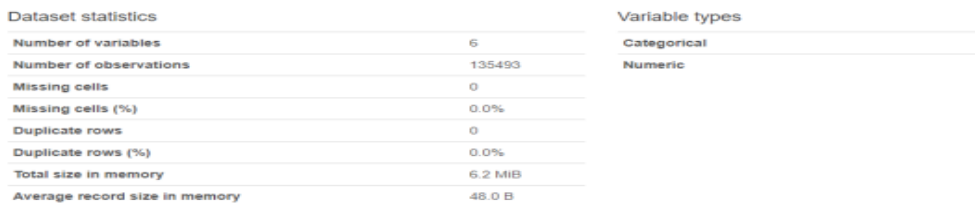
PROBLEM STATEMENT

The Tunisian Company of Electricity and Gas (STEG) is facing a problem of significant financial losses due to fraudulent manipulations of meters by consumers. Our aim is to develop a solution that can analyze the client's billing history data, identify patterns and anomalies that deviate from the norm, and flag them as potential fraud. The solution should be able to classify the billing history data into fraudulent and non-fraudulent activities with high accuracy. Additionally, the solution should be able to identify patterns and correlations in the data that may indicate fraudulent behavior and provide a visual representation of the data for better understanding and analysis. The ultimate goal is to help STEG to detect and prevent fraudulent activities, enhance revenues, and reduce financial losses.

DATASET DESCRIPTION

The data provided by STEG is composed of two files. The first one consists of client data and the second one contains billing history from 2005 to 2019.
There are 2 .zip files for download, train.zip, and test.zip and a SampleSubmission.csv. In each .zip file you will find a client and invoice file.

The below image describes the client data that is to be used for training the models



Dataset statistics

Number of variables	6
Number of observations	135493
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	6.2 MiB
Average record size in memory	48.0 B

Variable types

Categorical
Numeric

The below image describes the invoice data that is to be used for training the models

```
invoice.shape
```

```
(4476749, 16)
```

```
 #   Column                Dtype
---  ------                -----
 0   client_id             object
 1   invoice_date          object
 2   tarif_type            int64
 3   counter_number        int64
 4   counter_statue        object
 5   counter_code          int64
 6   reading_remarque      int64
 7   counter_coefficient   int64
 8   consommation_level_1  int64
 9   consommation_level_2  int64
 10  consommation_level_3  int64
 11  consommation_level_4  int64
 12  old_index             int64
 13  new_index             int64
 14  months_number         int64
 15  counter_type          object
```

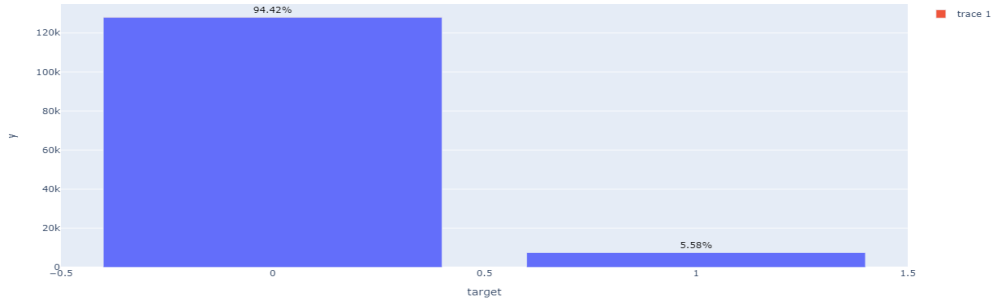
DATA PREPROCESSING

Initially the data sets were clean and had no null values, Some features required some data manipulation.
Extracted month, year from 'invoice_date', also added binary feature - 'is_weekday'
client_catg', 'district' and 'region' were assigned as categories to use them as categorical features in modeling.

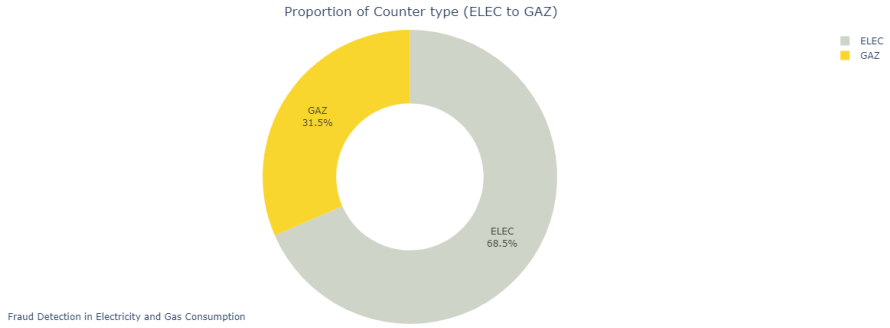
```
invoice_train['invoice_date']=pd.to_datetime(invoice_train['invoice_date'])
invoice_train["year"] = invoice_train["invoice_date"].dt.year
invoice_train["month"] = invoice_train["invoice_date"].dt.month
```

EXPLORATORY DATA ANALYSIS

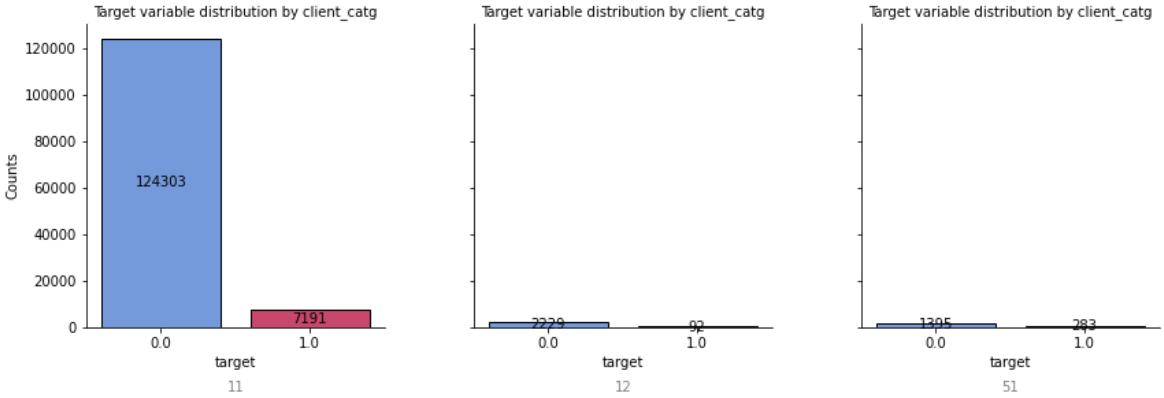
Below chart shows the distribution of target feature i.e 94.42% non fraudulent consumers and 5.58 % fraudulent customers.the dataset is highly imbalanced



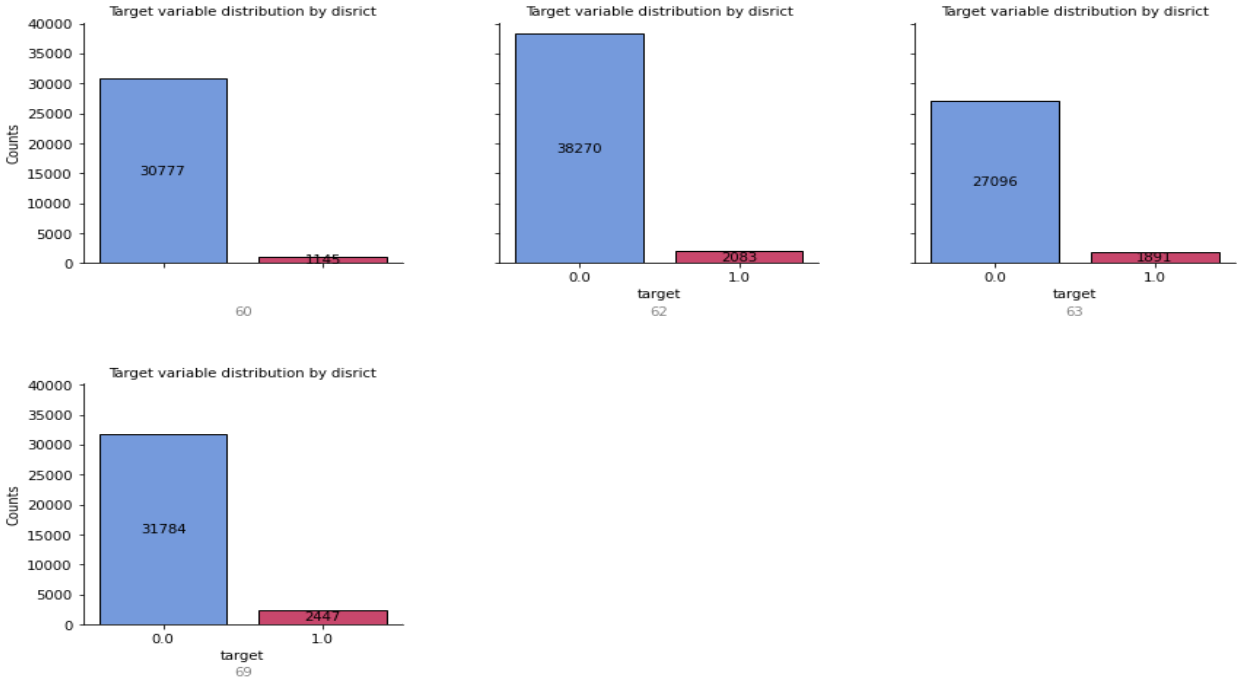
Below chart shows the distribution of gas and electricity consumers



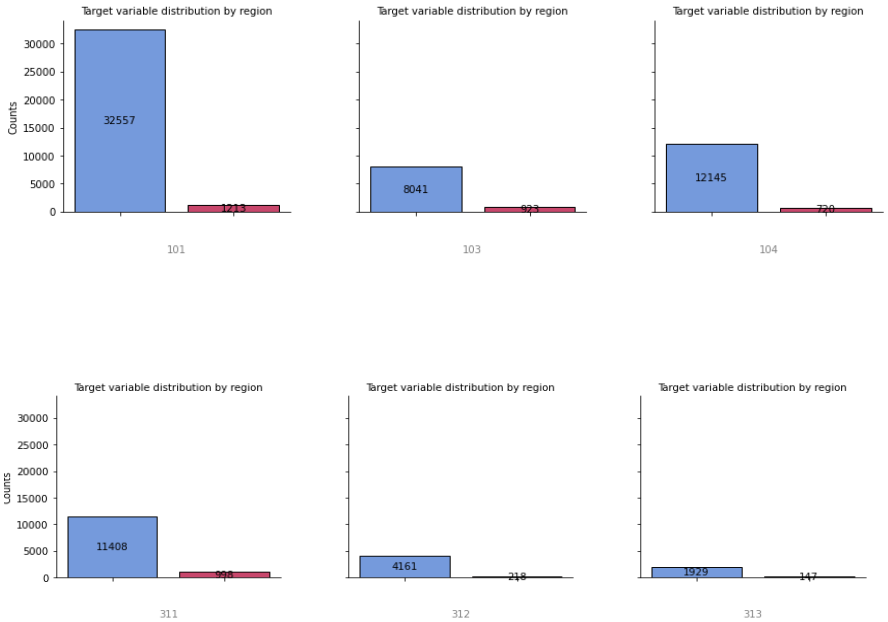
Below chart shows the distribution of consumers in different categories and we can see that the most populated category of consumers is labeled as 11



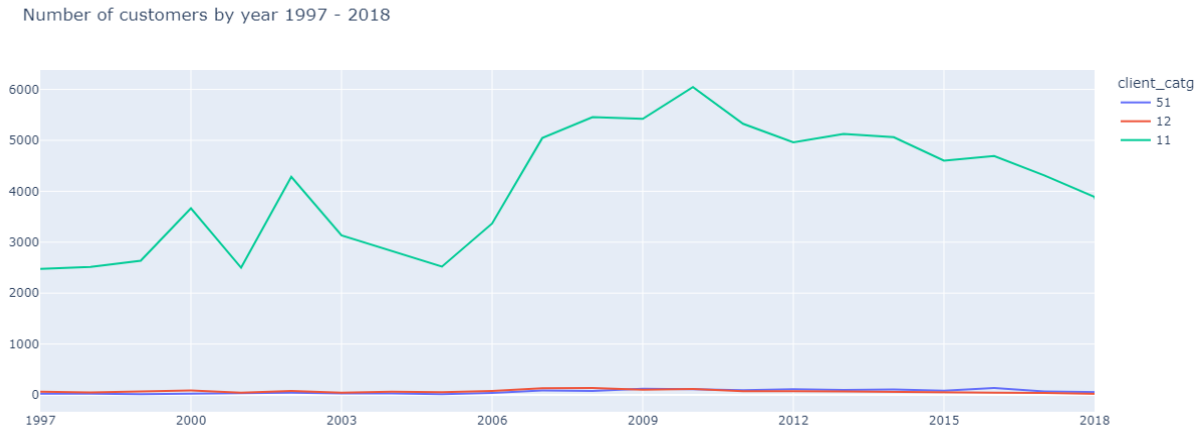
Below chart shows the distribution of consumers with respect to districts.



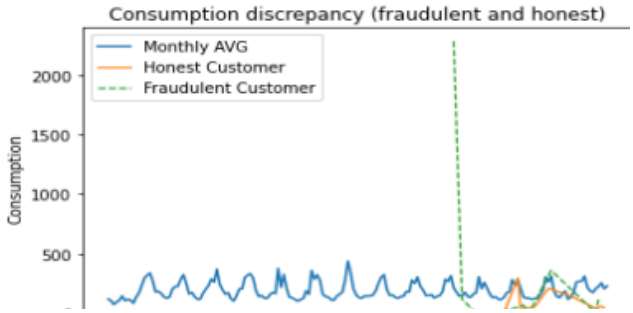
Below chart shows the distribution of consumers according to regions and we can see that in some regions the consumers are very low this may be due to the fact that the consumers in those regions are having electric or gas supply from other company. Or may be the data is not available from those regions.

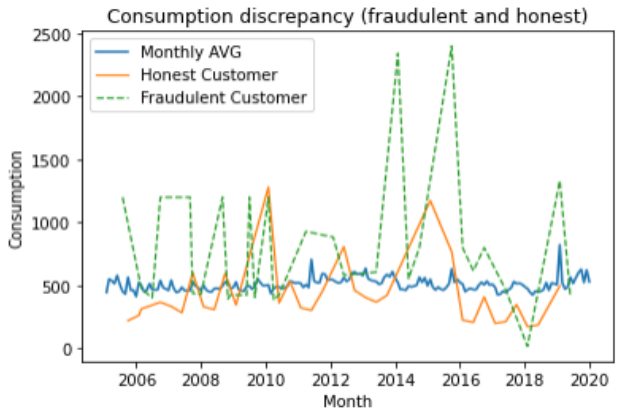


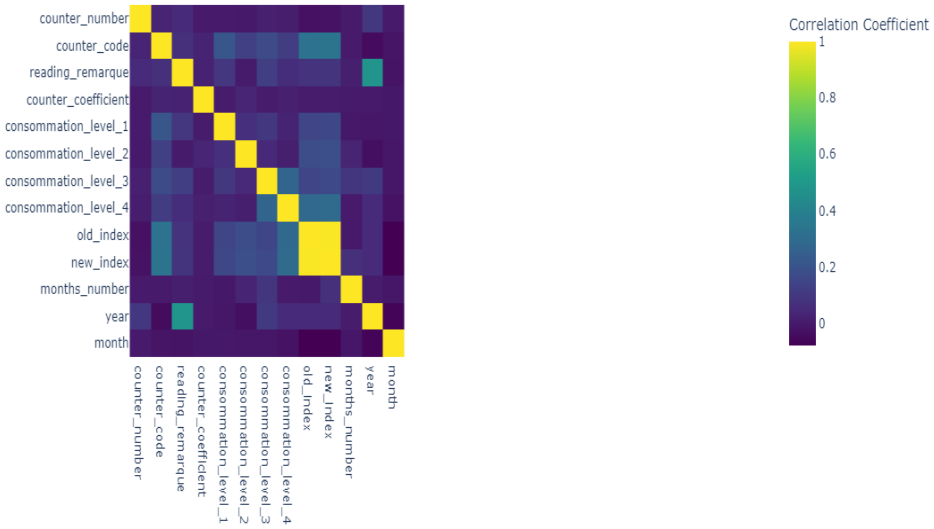
Below graph shows the number of consumers joining the company from 1997-2018 and the peak year we have is 2010 in which the most number of consumers joined the company.



Below graph shows the discrepancy between the honest and dishonest clients by showing the monthly avg consumption and comparing it with the avg consumption of an honest and dishonest client. We randomly checked different clients and we were able to see the prominent differences between honest and dishonest clients.







Relatively some correlations were also found between some features in the invoice dataset

FEATURE ENGINEERING

The features of electric and gas consumption levels, old and new index, tariff type, and counter status are essential in creating new features for fraud detection in electricity and gas consumption. By analyzing patterns and deviations in these features, it is possible to identify irregularities in consumption that may indicate fraudulent behavior. Additional features such as monthly consumption levels, reading remarks, and invoice dates can also be used to improve the accuracy of fraud detection algorithms. With the help of machine learning classification techniques and feature engineering, it is possible to create effective models that can accurately detect instances of fraud and reduce losses for utility companies.

Features that were used to generate new features

- Electric and Gas Consommation levels (1-4)
- Old Index
- New Index
- Tariff Type
- Counter Status

METHODS USED IN FEATURE ENGINEERING

Feature engineering is a crucial step in developing accurate models for fraud detection in electricity and gas consumption. Various statistical methods such as cumulative sum, measures of central tendencies (mean, mode, median), and measures of spread (range, standard deviation, variance) can be used to generate new features from existing data.

For instance, cumulative sum techniques can help identify trends and patterns in consumption data and highlight any deviations from the expected trend that may indicate fraudulent activity. Measures of central tendencies such as the mean, mode, and median can provide insights into the typical consumption behavior of customers and help identify outliers. Similarly, measures of spread such as the range, standard deviation, and variance can provide information about the variability in consumption patterns and highlight any unusual patterns.

Group by methods in Pandas can be used to group the consumption data based on specific attributes such as consumer ID, tariff type, and counter status. This can help in identifying consumption patterns for specific groups of customers and identifying any unusual behavior within those groups. By combining these methods and creating new features based on the insights gained, it is possible to improve the accuracy of fraud detection models and reduce losses for utility companies.
Below code snippet shows the function that we created for making new features.

```
summary_invoice_train = (
    invoice_train_cumsum.loc[:, ~invoice_train_cumsum.columns.isin(["counter_code", "counter_number"])].groupby(["client_id", "counter_type"]).agg(avg_consom_l_1=("consommation_level_1", "mean"),
        var_consom_l_1=("consommation_level_1", "var"),
        sd_consom_l_1=("consommation_level_1", "std"),
        median_consom_l_1=("consommation_level_1", "median"),
        mode_consom_l_1=("consommation_level_1", find_mode),
        avg_diff_consom_l_1=("consommation_level_1", lambda x: np.mean(np.diff(x))),
        range_consom_l_1=("consommation_level_1", lambda x: np.max(x) - np.min(x)),
        sd_cumsum_consommation_level_1=("cumsum_consommation_level_1", "std"),
        avg_cumsum_consommation_level_1=("cumsum_consommation_level_1", "mean"),
```

BUILDING MACHINE LEARNING MODELS

Based on the results of our baseline ML models, it appears that the best-performing model in terms of overall accuracy is XG Boost, with an accuracy score of 0.944. However, accuracy is not always the best metric to evaluate the performance of a classification model, especially in imbalanced datasets such as yours. It's also important to consider precision, recall, and AUC when evaluating your models.

Looking at precision, which measures the proportion of true positives among all positive predictions, LGBM Boost seems to perform the best with a score of 0.967. However, its recall

score of 0.4 indicates that it may struggle with identifying all of the positive cases in your dataset.

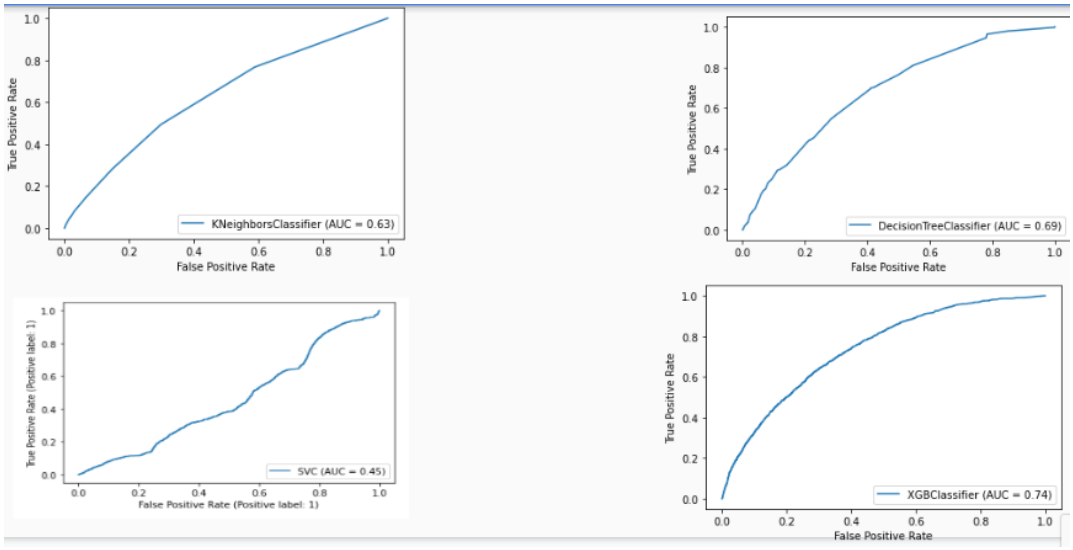
AUC, which measures the overall ability of a model to distinguish between positive and negative cases, also provides useful information about model performance. Here, XG Boost seems to be the best-performing model with an AUC score of 0.74.

In summary, while XG Boost may have the highest overall accuracy, LGBM Boost may be a better choice if precision is a more important metric for your use case. However, it's important to also consider other metrics such as recall and AUC when selecting the best model for your specific needs.

Below table shows the initial results of models without feature engineering involved

MODELS	ACCURACY	PRECISION	RECALL	AUC
Linear Reg	0.943	0.14	0.0568	067
Decision Tree	0.891	0.007	0.002	0.69
KNN	0.940	0.000	0.000	0.63
SVM	0.943	0.000	0.000	0.45
XG BOOST	0.944	0.22	0.002	0.74
ADA BOOST	0.943	0	0	0.69
CAT BOOST	0.944	0.29	0.017	0.72
LGBM BOOST	0.941	0.967	0.4	0.73

Below are the resultant ROC curves for the initial models



In the above figure we can see that xgb boost is giving us promising results and on the other hand SVM is the worst.

MODEL RESULTS AFTER USING THE UPDATED DATASET

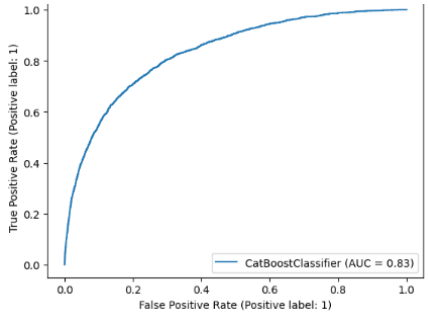
After watching the initial performance of the models we decided to work on boosting models only that are XGB boost, Ada boost, Catboost and lgbm boost. Because these are the models that are giving us better performance and we can work on these to improve the performance. Below figure shows the results of these boosting models after we trained these models on 137 features.

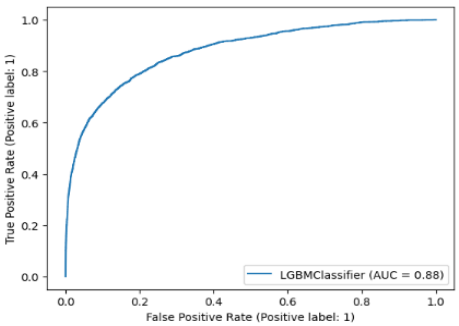
MODELS	ACCURACY	PRECISION	RECALL	AUC
XG BOOST	0.936	0.48	0.07	0.81
ADA BOOST	0.934	0	0	0.76
CAT BOOST	0.938	0.67	0.09	0.83
LGBM BOOST	0.961	0.95	0.4	0.88

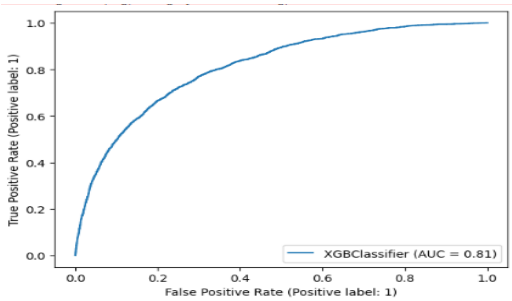
Before feature engineering, the highest evaluation metric was achieved by LightGBM Boost with a score of 0.941, followed by XGBoost with a score of 0.944, CatBoost with a score of 0.938, and AdaBoost with a score of 0.943.

After feature engineering, LightGBM Boost still achieved the highest evaluation metric with a score of 0.961. However, there was a significant improvement in the metrics of the other algorithms.
Overall, it seems that feature engineering had a positive impact on the performance of these algorithms, particularly for XGBoost and lgbm Boost.

Below figure shows the ROC curve of the models and we have observed that there is significant improvement in the models







DATA BALANCING TECHNIQUES

We have used oversampling and undersampling techniques to observe how the models are behaving because in a fraud detection project, it is important to choose the appropriate balancing technique based on the specific characteristics of the dataset and the performance of the model. It is also important to evaluate the effectiveness of the technique and ensure that the model is not biased towards the majority class or overfitting on the minority class.

MODEL PERFORMANCES AFTER OVERSAMPLING

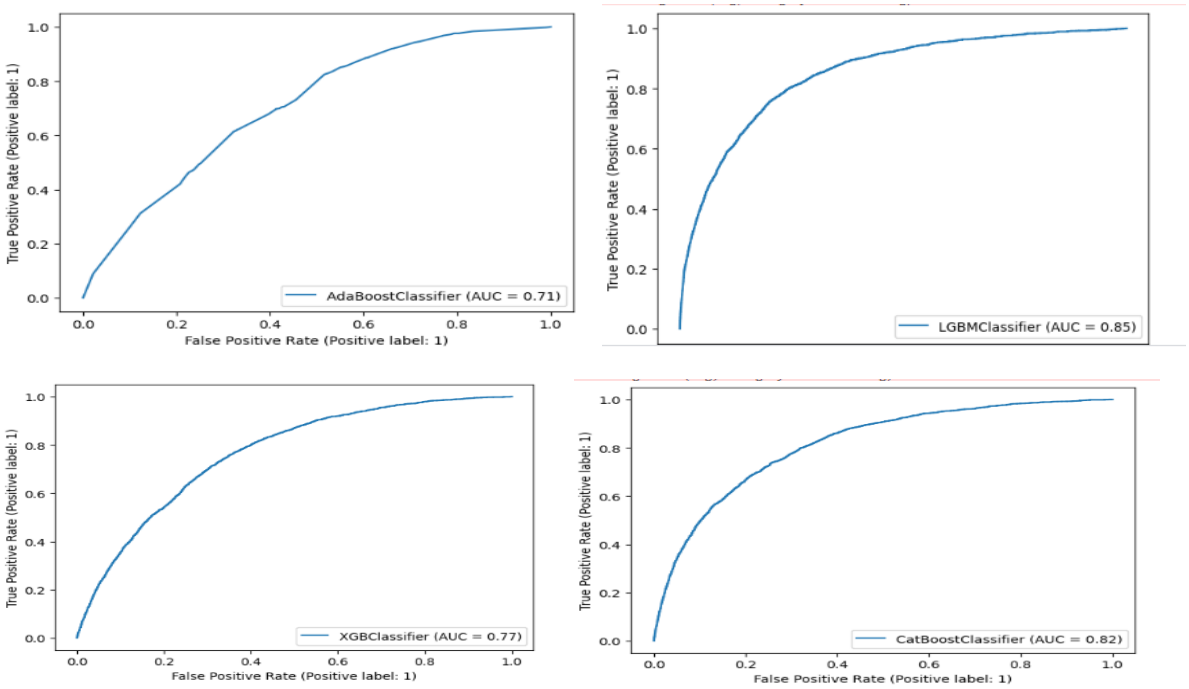
Below figure shows the results after performing oversampling:

MODELS	ACCURACY	PRECISION	RECALL	AUC
XG BOOST	0.844	0.17	0.42	0.74
ADA BOOST	0.92	0.21	0.08	0.71
CAT BOOST	0.92	0.37	0.17	0.82
LGBM BOOST	0.87	0.27	0.58	0.85

After applying over-sampling, it appears that the accuracy has generally decreased for all models, which can be expected when balancing the dataset by oversampling the minority class. However, it is important to note that accuracy is not always the most informative metric, especially for imbalanced datasets.

The precision metric, which measures the proportion of true positives among all positive predictions, has improved for all models, which suggests that the models are better at correctly identifying the minority class. The recall metric, which measures the proportion of true positives among all actual positives, has also improved for most models, indicating that the models are better at identifying more of the actual fraudulent transactions.

The AUC (Area Under the Curve) metric, which measures the overall performance of the model across different probability thresholds, has improved or remained similar for all models except AdaBoost, which has decreased slightly.



Overall, it appears that the over-sampling technique has led to improvements in precision and recall for all models, which are important metrics for fraud detection. However, the accuracy has decreased, which suggests that the models are making more errors in classifying the majority class. It is important to carefully evaluate the trade-offs and choose the appropriate balancing technique based on the specific characteristics of the dataset and the performance of the model.

MODEL PERFORMANCES AFTER UNDERSAMPLING

Results after undersampling:

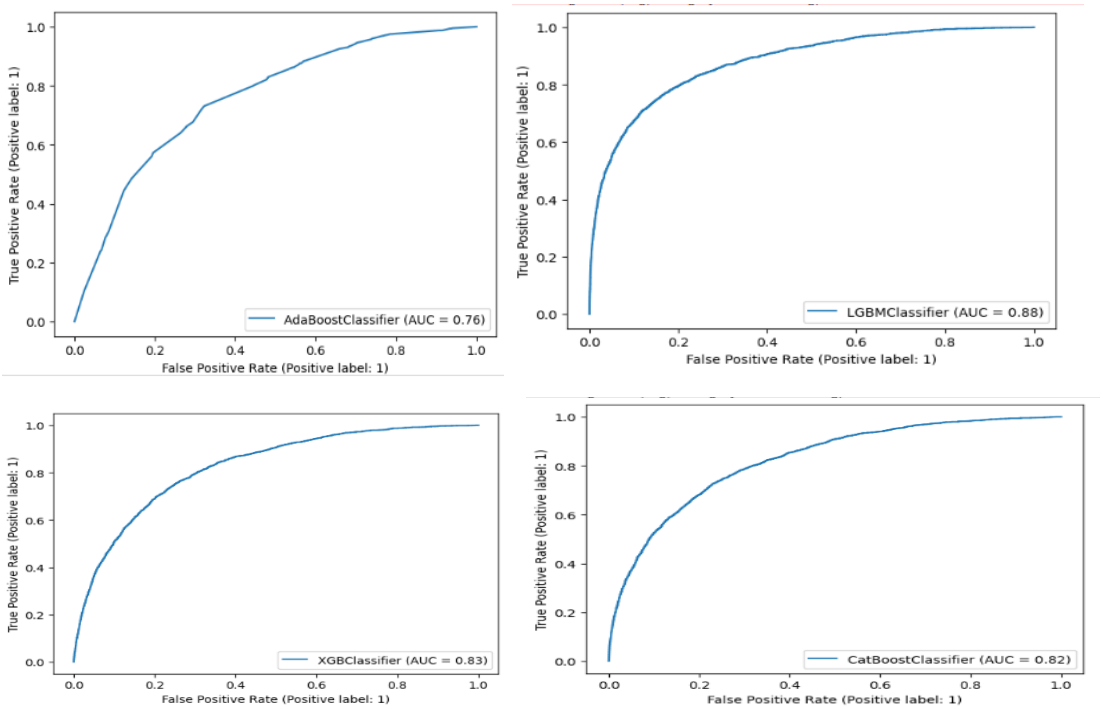
MODELS	ACCURACY	PRECISION	RECALL	AUC
XG BOOST	0.93	0.42	0.12	0.83
ADA BOOST	0.93	0.00	0.00	0.76
CAT BOOST	0.938	0.64	0.09	0.82
LGBM BOOST	0.94	0.88	0.4	0.88

After applying under-sampling, the accuracy has remained similar or increased slightly for all models. However, it is important to note that accuracy is not always the most informative metric, especially for imbalanced datasets.

The precision metric, which measures the proportion of true positives among all positive predictions, has varied across the different models. XGBoost and LightGBM Boost have the highest precision scores, indicating that they are better at correctly identifying fraudulent transactions. AdaBoost has a precision score of 0, which means it did not identify any true positives, suggesting that the model may not be suitable for this task.

The recall metric, which measures the proportion of true positives among all actual positives, has also varied across the different models. XGBoost and LightGBM Boost have low recall scores, indicating that they are not identifying as many of the actual fraudulent transactions. CatBoost and AdaBoost have recall scores of 0, indicating that they did not identify any of the actual fraudulent transactions.

The AUC (Area Under the Curve) metric, which measures the overall performance of the model across different probability thresholds, has remained similar or decreased slightly for all models except LightGBM Boost, which has remained the same.



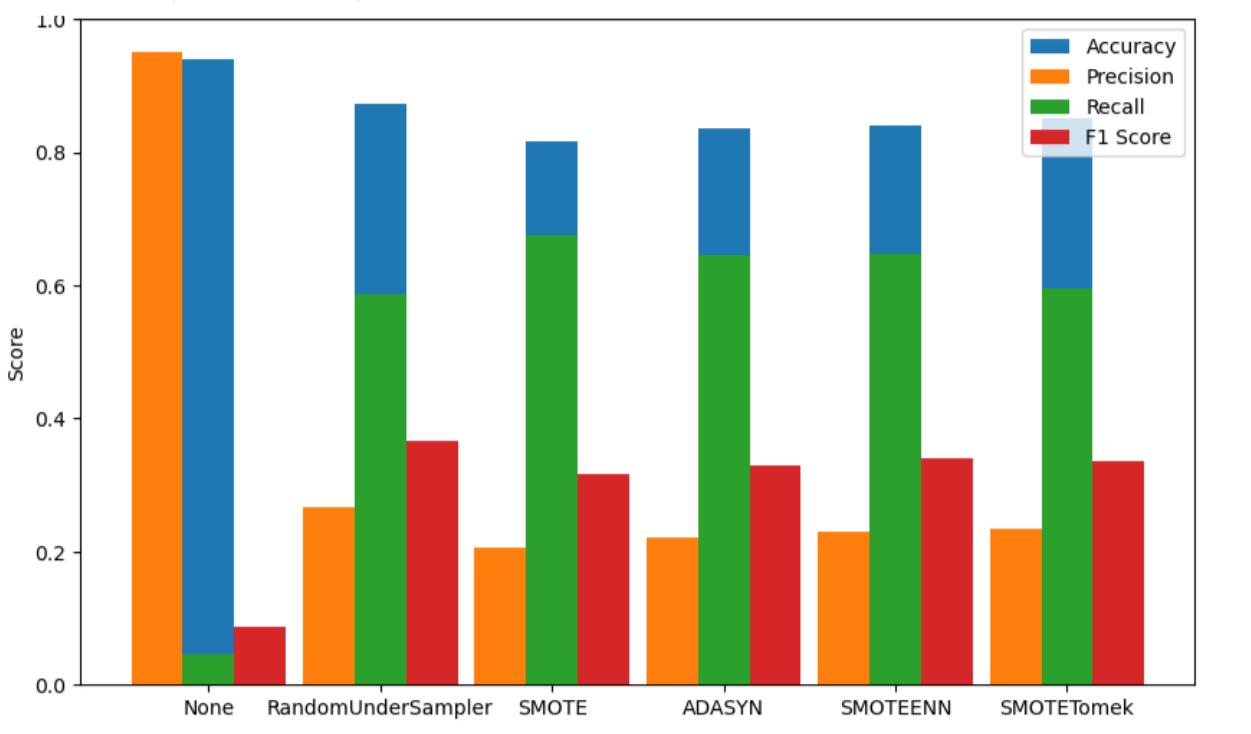
Overall, it appears that the under-sampling technique has led to improvements in precision for some models, but has negatively affected recall for others. It is important to carefully evaluate the trade-offs and choose the appropriate balancing technique based on the specific characteristics of the dataset and the performance of the model.

FURTHER INVESTIGATION

We further made an investigation to see the effects of different types of sampling techniques on the model.The results are shown on the diagrams below.

Comparison and checking the effects of different other sampling techniques on LGBM Classifier

Metric Sampling Technique	Accuracy	F1 Score	Precision	Recall
None	0.94	0.087	0.95	0.046
RandomUnderSampler	0.87	0.37	0.27	0.59
SMOTETomek	0.85	0.34	0.23	0.6
SMOTEENN	0.84	0.34	0.23	0.65
ADASYN	0.84	0.33	0.22	0.64
SMOTE	0.82	0.32	0.21	0.68



We have noticed that different sampling techniques did not yield better results.We therefore stick to our original model.

FUTURE WORK

Feature engineering: Feature engineering is the process of selecting and creating features that have a significant impact on the model's predictive power. We can explore more data sources or create new features based on existing data to improve the model's accuracy and feature selection techniques can be utilized to find best features.

Hyperparameter tuning: Hyperparameters control the learning process of the model. We can do more experiments with different hyperparameters of the LGBM model to optimize its performance further.

Ensemble models: We can also explore the effectiveness of combining multiple models using ensemble techniques such as bagging or boosting to improve the overall performance.

Deep Learning models: Deep learning models such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) have shown promising results in fraud detection. We can consider exploring these models to improve the detection rate.

Interpretability: While achieving high accuracy is important, it is also crucial to understand how the model is making decisions. We can explore ways to interpret the model's output and provide insights into the key factors that contribute to fraud detection.

CHALLENGES

As the most prominent challenge for us was the resources available. The points that we have mentioned as future work were not being implemented because of the lack of CPU powers. Since the data was very big to handle in our local machines and even the cloud services like google colab and kaggle were crashing when we were trying to implement the techniques we have mentioned above..

CONCLUSION

Our analysis has shown that the LGBM model is the most effective model in detecting fraud, with an accuracy of 94 percent, precision of 96 percent, and an AUC score of 91. Although the recall rate of 8 percent is relatively low, it is still considered satisfactory as the cost of false negatives is much higher than false positives.Our analysis has shown that the LGBM model is the most effective model in detecting fraud, with an accuracy of 94 percent, precision of 96 percent, and an AUC score of 91. Although the recall rate of 8 percent is relatively low, it is still considered satisfactory as the cost of false negatives is much higher than false positives.

REFERENCES

1.Machine Learning Yearning,Technical strategy for AI engineers in the era of deep learning (Andrew NG)
2.PATTERNS, PREDICTIONS, AND ACTIONS A story about machine learning (Moritz Hardt and Benjamin Recht)
3.Probability & Statistics for Engineers & Scientists(Ronald E. Walpole Roanoke College Raymond H. Myers Virginia Tech Sharon L. Myers Radford University Keying Ye University of Texas at San Antonio)
4.An Introduction to Decision Modeling with DMN.
5.Taylor, James (2011). Decision Management Systems – A Practical Guide to Using Business Rules and 6.Predictive Analytics. IBM Press. DeBevoise, Tom and Taylor, James (2014). The MicroGuide to Process and Decision Modeling with BPMN/DMN.