Data Mining Project 2023-2024

Credit Card Fraud Prediction Dataset Analysis

by

Anna Karolin

Université Jean Monnet
Saint-Étienne

Data Mining Project 2023-2024

Credit Card Fraud Prediction Dataset Analysis

---

1. Introduction

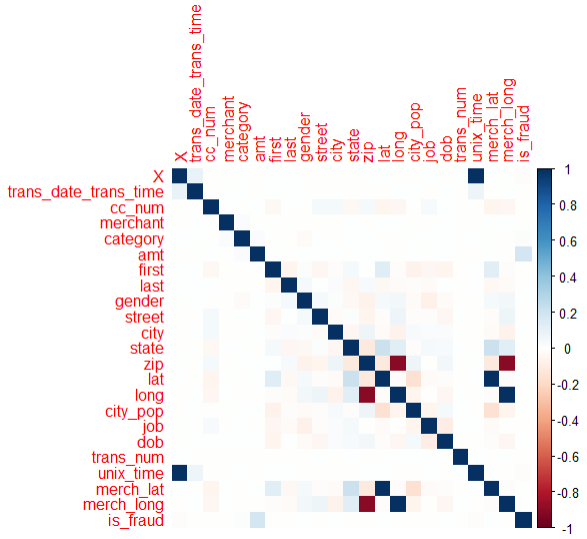
Classification models are supervised machine learning models that predicts the correct label of a given input data. Classification tasks can either be binary classification or multi-class classification. Some of the most commonly used classification models are decision trees, logistic regression, random forests, neural networks etc. In this project I explore the some of these classification models on the Credit Card Fraud Prediction Dataset to make an analysis of the models based on various accuracy measures.

2. Presentation of the Dataset

The dataset we chose to do the analysis is the Credit Card Fraud Prediction Dataset `https://www.kaggle.com/datasets/kelvinkelue/credit-card-fraud-prediction`. The dataset was created by Kelvin Kelue on kaggle and is available by clicking on the link provided.

The dataset consists of 555719 observations with 23 features each. The dataset is for binary classification. All the features are of numeric, string or alpha-numeric data type and the target variable is of data type factor. The dataset contains details of transaction such as the name of the transactor, their occupation, reason for the transaction, the latitude and longitudes of the transactor, latitude and longitude of the merchant, various details of the merchant and so on. Due to the large size of the data, the models are trained on a random sample keeping the percentage of the two classes same.

The two classes in the dataset are highly imbalanced and the binary classification deals with the transaction is not fraud (class 0) and the transaction is fraud (class 1). The dataset is almost balanced with 99.61% of class 0 and 0.3860% of class 1. There are also no missing values present in the dataset.

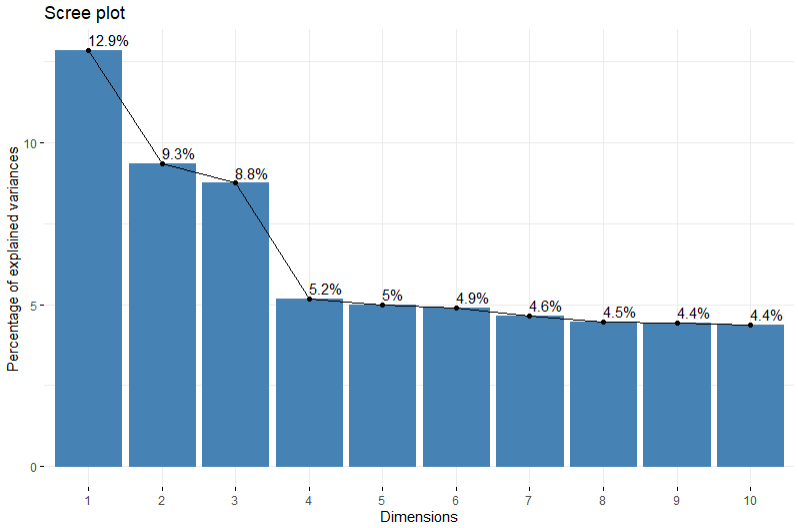


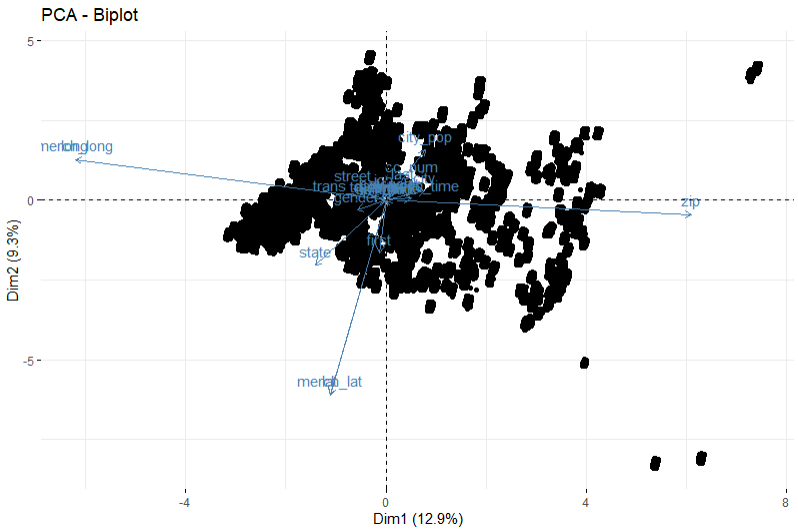
The correlation map of the features are as above. As seen in the map, there are not many features with high correlation. Most of the features are uncorrelated to having very little correlation. But the available correlation helps us getting a general idea to the feature which helps most with the classification task at hand.

3. Preparation of the Dataset

The first step in processing the data was looking for any missing values, which for he dataset turned out to be that there are no missing values. Then all the features were converted to numeric for obtaining the correlation plot and training the models.

Principal Component Analysis (PCA) reduces the dimensionality of the data for visualisation, exploratory analysis and pre-processing. PCA was applied to the data and the following graphs showing importance of each new feature created was obtained.





4. Classification Models

Due to the large size of the data, all the classification models trained were done on a random sample taken from the data. Due to the heavy imbalancy present, while taking the random sample the proportion of data points from id each class is kept same as their proportion in the entire data set. The accuracy metrics used for the comparison of the trained models are confusion matrix and the f-score as the accuracy of the models will be biased due to the heavy imbalancy. The following classification models were trained in this project:

- Logistic Regression Model
- Decision Tree Model
- Support Vector Machines
- Random Forest Model

4.1. Logistic Regression

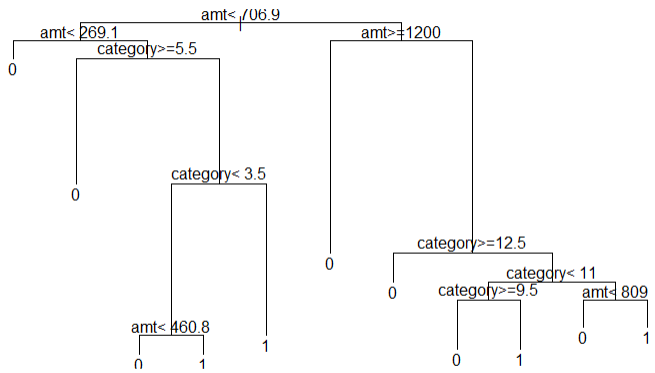
Logistic Regression is a supervised classification model which predicts the probability that an input belongs to which of the classes. The confusion matrix after training the models and testing it on a test set is as follows:

		Predicted	
		Negative	*Positive*
Actual	*Negative*	33176	300
	Positive	38	0

The accuracy is 0.99 , the precision and recall values are 0 and the f-scores cannot be calculated. We can observe from the confusion matrix itself that the classifier is not doing a good job.

4.2. Decision Trees

Decision trees are classification models that takes a tree structure commonly used for classification and regression tasks. It can train the data even when the features are of different data types and works well with discrete features. For uniformity, I have trained the decision tree with all the features converted to numerical type. The tree obtained is as follows:



The confusion matrix after training the models and testing it on a test set is as follows:

		Predicted	
		Negative	*Positive*
Actual	*Negative*	33178	125
	Positive	36	175

The accuracy is 0.99 , the precision and recall values are 0.5833 and 0.8294 respectively and the f1, f2 and f3 scores are 0.6849, 0.7649, 0.7958 respectively.

4.3. Random Forest

Random Forests combine the output of several decision trees to give an optimal prediction.

4.3.1. $n = 50$

In this model I trained, I chose the number of trees, n, to be 50. The confusion matrix after training the models and testing it on a test set is as follows:

		Predicted	
		Negative	*Positive*
Actual	*Negative*	33198	108
	Positive	16	192

The accuracy is 0.99 , the precision and recall values are 0.64 and 0.9231 respectively and the f1, f2 and f3 scores are 0.7559, 0.8481, 0.8840 respectively.

4.3.2. $n = 100$

In this model I trained, I chose the number of trees, n, to be 100. The confusion matrix after training the models and testing it on a test set is as follows:

		Predicted	
		Negative	*Positive*
Actual	*Negative*	33203	109
	Positive	11	191

The accuracy is 0.99 , the precision and recall values are 0.6367 and 0.9455 respectively and the f1, f2 and f3 scores are 0.7610, 0.8619, 0.9018 respectively.

4.4. Support Vector Machines

Support Vector Machines are supervised models primarily used for the tasks of classification and regression. In classification task SVMs tries to give a good classifier by increasing the margin of the classifier with the help of support vectors. The confusion matrix after training the models and testing it on a test set is as follows:

		Predicted	
		Negative	*Positive*
Actual	*Negative*	33214	300
	Positive	0	0

The accuracy is 0.99 , the precision and recall values are 0 and the f-scores cannot be calculated. We can observe from the confusion matrix itself that the classifier is not doing a good job.

5. Observation and Analysis

Among the models trained we can observe that random forest models are classifying the data better with the f1, f2 and f3 scores are 0.7559, 0.8481, 0.8840 and 0.7610, 0.8619, 0.9018 respectively for number of trees = 50 and 100. On the other-hand, we can also see that logistic regression model and support vector machines completely fail in the classification task with not identifying a single fraudulent transaction as fraud. Also the decision boundary plots are not included in the report or the code part as the two classes are heavily mixed and the high imbalancy in the classes makes the decision boundary not to purpose.

6. Conclusion

The data being highly imbalanced affects the performance of the models. Support vector machines are generally considered to be one of the good classifiers but with this dataset dataset, it's not working as good because of the imbalancy and the distribution of the classes. The next step that can be done to improve the models using hyperparameter tuning.

The reason for the choice of the problem of classification models was to understand how the models work. In addition to that the high class imbalancy also affected the performance of the models.

7. Reference

- `https://www.kaggle.com/datasets/kelvinkelue/credit-card-fraud-prediction`
- `https://scikit-learn.org/stable/modules/tree.html`
- `https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/`

8. Appendix

8.1. Pairplots of Difficult Features

Here the blue points indicates transactions which are not fraudulent while orange points indicates fraudulent transactions. We can see that both the classes are non separable on graph and hence I avoided plotting decision boundaries (The below pairplot is generated using python for better visualization).