SAPIENZA
UNIVERSITÀ DI ROMA

Department of Information Engineering, Informatics and Statistics

Fraudolent Transaction Detection

Foundamentals of Data Science

Professors:
Professor Fabio Galasso

Students:
Giuseppe Di Poce
Angelo Mandara
Claudiu Gheorghiu
Enrico Grimaldi
Tito Tamburini

Academic Year 2022/2023

1 Abstract

We are going to develop a model for predicting fraudulent transactions for a financial company and use insights from the model to develop an actionable plan. We start by pre-processing the data so as to clean the data set, transform each feature into a numeric type, and remove uninformative data, i.e. reduce the complexity of the problem. Then we proceed with the selection of the best model in association with certain sampling techniques, in particular we choose to apply a decision tree model on an oversampled data set.

2 Introduction

The problem reported is of binary classification. We therefore report an extremely current issue given that digital transactions are exponentially increasing, so security in this context is critical. In this way is possible to increase the user base for this type of service, boosting the traceability of payments, control of shady money movements and reducing the tax evasion rate. To pursue an acceptable classification of our examples we compare a number of unsupervised methods, selecting the best model. In particular we focus on Logistic Regression and Decision Tree, and then try to implement an Xgboost algorithm based on a Random Forest Classifier. The data set chosen by Kaggle has ten features and about six and a half million observation examples.

3 Related work

(a) https://www.kaggle.com/code/sasakitetsuya/fraudulent-transactions-analysis-by-xgboost

(b) https://www.kaggle.com/code/dibyendupatra/fraudulent- transactions-data

(c) https://www.kaggle.com/code/kartik2khandelwal/predicting-fraudulent-transactions

4 Data set and pre-processing

The data set has as its main challenges the presence of nonnumeric variables and a strong output variable imbalance. We addressed these challenges by exploiting hashing techniques for strings, one-hot encoding for categorical features, and sampling and multi-collinearity analysis techniques to generally improve performance (consider the notebook and explanation below). Finally, we delved into dimensionality reduction by dropping non-informative columns (the card types that do not contain fraudulent transactions and the fraudulent transaction alert).

5 Proposed methods and results

5.1 Multi-Collinearity Analysis

In statistics, multicollinearity is a phenomenon in which one predictor variable in a multiple regression model can be linearly predicted from the others with a degree of accuracy. The coefficient estimates of the multiple regression may change erratically in response to small changes in the model or the data. Multicollinearity does not reduce the predictive power or reliability of the model as a whole; it only affects calculations regarding individual predictors. "No multicollinearity" refers to an exact (non-stochastic) linear relation among the predictors. A formal detection-tolerance for the variance inflation factor (VIF) for multicollinearity is given by:

$$tolerance = 1 - R_j^2 \,, \quad VIF = \frac{1}{tolerance}$$

where R_j^2 is the "coefficient of determination" of a regression of explanator j on all the other explanators. A tolerance of less than 0.20 or 0.10 and/or a VIF of 5 or 10 and above indicates a multicollinearity problem. We solved this point collapsing four critical columns in two columns more informative and less redundant.

5.2 Logistic Regression

It's a technique for forecasting a categorized outcome variable from a set of individual factors. The Sigmoid function reflects the probability of items to be classified as fraudulent or not, i.e. decision boundary:

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$

We note that the quality of the model in our case (in terms accuracy and precision - Fig.2) is low. However, we prefer to abandon a logistic classification in favor of a Decision tree model. In fact, as we can see the Logistic Regression and Decision trees differ in the way that they generate decision boundaries (Fig. 1): Decision Trees bisect the space into smaller and smaller regions, whereas Logistic Regression fits a single hyperplane. Indeed we obtained as results accuracy and recall of 0.50 and a precision approximated to zero.

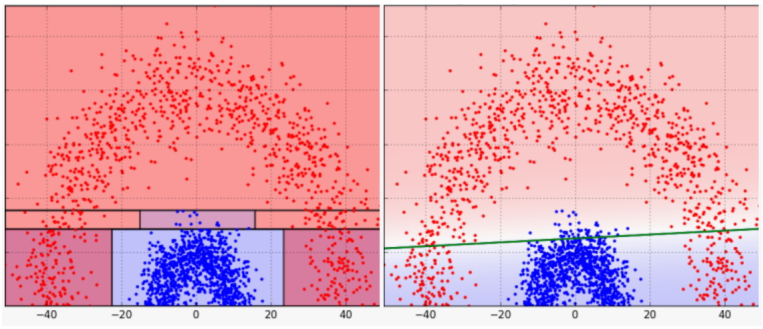

Figure 1: Decision boundaries, Logistic vs Decision Tree

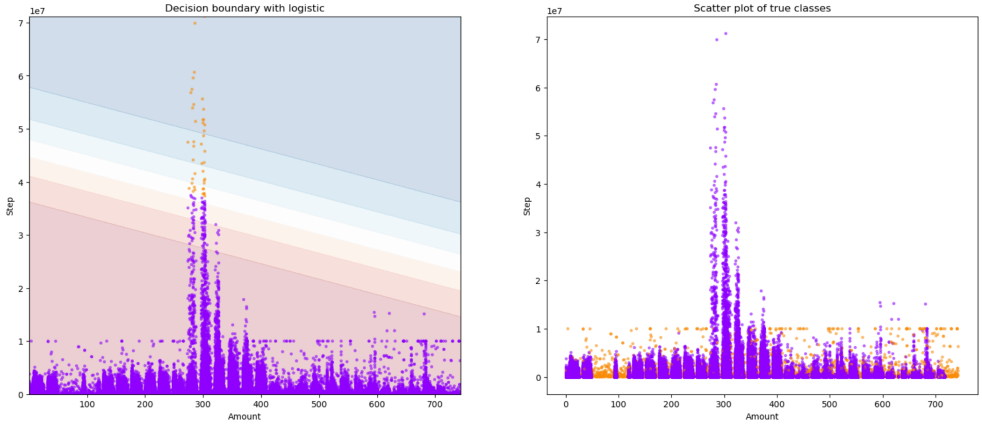

Figure 2: Logisitc Regression performance

5.3 Sampling techinques with Decision Tree

A Decision Tree (DT) uses a tree structure to specify sequences of decisions and consequences. Given the set of input variables $X = \{X_1, X_2, ...X_n\}$ the goal is to predict a response or output variable . A decision tree can be converted into a set of decision rules and can be applied to a variety of situations. The following resampling techniques will be useful in making the data set balanced according to the number of fraudulent and non-fraudulent transactions. Among the techniques used the undersampling one is to consider a smaller number of the minority class. Oversampling, on the other hand, involves balancing the data set based on the redundancy of examples of the training set belonging to the minority class; finally, SMOTE is a hybrid of the previous two methods. The obtained results (Fig. 3) highlight how the oversampling technique optimizes the trade-off between precision and recall leading to the following metrics: $Precision = 0.78$, $Recall = 0.72$, $F1score = 0.75$.

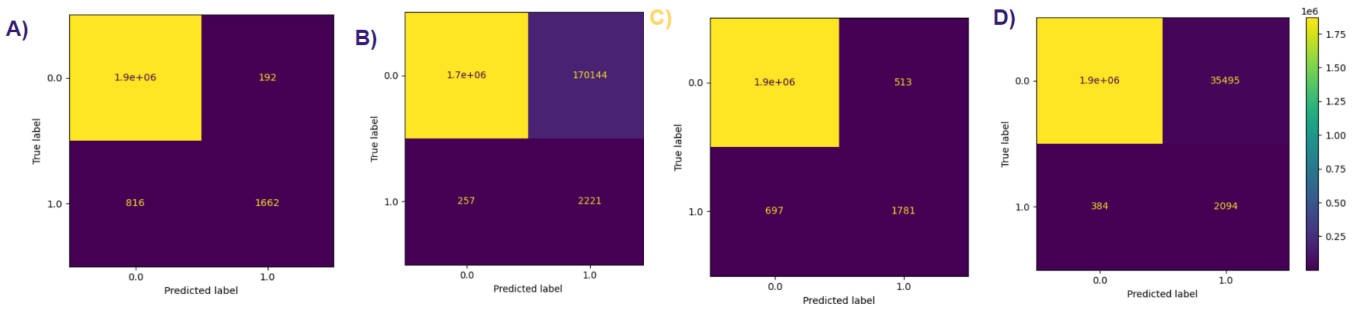

Figure 3: Confusion matrices according to the various sampling techniques:
A) without resampling, B) undersampling, C) oversampling, D) SMOTE.

5.4 Random Forest and Xgboost

Random forests consist in building an ensemble of decision trees grown from a randomized variant of the tree induction algorithm. Decision trees are indeed ideal candidates for ensemble methods since they usually have low bias and high variance, making them very likely to benefit from the averaging process. We also decide to implement a Random Forest Classifier with Xgboost making the process more performant: it consists of a sequential choice of decision trees based on reducing the error of the previous tree. However, the results obtained are identical to those from simple Decision Tree associated with oversampling of fraudulent transactions, probably due to the various arrangements made in the pre-processing phase.

5.5 Conclusions

A fundamental point for us (and an eventual company) is to concentrate our efforts on predictive models that minimize FALSE NEGATIVES and we would increase the recall but without losing too much efficiency in terms of precision. The accuracy level was arleady high before we applied sampling techniques because of the umbalanceness of the data set. We also have some suggestions for the interested companies. Gathering more information about time of a transaction 'step' feature is important according to our analysis but is not really informative given the fact that we don't know the 'zero time' and we can't extract further information. Then they should pay more attention on the 'transfer' type cards in combination with low amounts of money transferred.

6 References

[ ] https://www.kaggle.com/code/kartik2khandelwal/predicting-fraudulent-transactions

[ ] Pattern recognition and machine learning, C. Bishop

[ ] https://quantifyinghealth.com/vif-threshold/

7 Repository with plots and code

https://github.com/sosamandara/FDS_Final_Project