Credit Card Fraud Detection using Classification

Team ID: 10
Name: Jordan Brennan ; Exploratory Data Analysis & Visuals, Random Forests
Name: Kideok Kwon ; Modeling, Prediction, Subset Selection, Automation
Name: Yuyan Li ; Background, Performance Metrics, Class Imbalance
Name: Samantha Soendoro ; Presentation Design, Sampling Methods, Final Words

Note: most topics were discussed, brainstormed, and written together. The assignments are based on the emphasis provided by each teammate

1 Introduction

This document is the final report for the course project of Group 10 of STA 141A, Fall 2019.

1.1 Background

In this project, we analyze **Credit Card Transactions** from an anonymous European Cardholder. This dataset contains 284,807 rows, with 31 features, one being the target variable, which is a binary variable that indicates if the transaction is Fraudulent or not. This is the variable we are interested in predicting.

As with many fraud-related data, this dataset is highly imbalanced, as there are much fewer fraudulent transactions than non-fraudulent transactions. More precisely, **out of the 284,807 rows, only 0.17% of them are Fraudulent**. Meaning, there are 485 fraudulent cases, with 284,322 legitimate transactions. Additionally, as this type of data is naturally confidential, the variables have been made anonymous. To make this dataset anonymous, PCA was applied to reduce dimensionality (from an unspecified number of original columns), which also has a bonus effect of hiding column names. (This was done by the data provider)

Table.1 Description of features

Features	**Description**
Time	Number of seconds elapsed between this transaction and the first transaction in the dataset
V1, V2, … , V28	Result of a PCA Dimensionality reduction to protect user identities and sensitive features
Amount	Transaction amount

Class	1 for fraudulent transactions, 0 otherwise

Due to the specified nature of the dataset, analyses must rely fully on statistical and machine learning theory. Domain knowledge on security will not be useful. This is an incredible project as it exposes the users to different sampling techniques and the importance of it, as well as understanding the trade-off between accuracy and computational speed between different machine learning algorithms and its parameters. Alongside that, this project is a great introduction to more modern data science trends in, as a dataset of this size is, without a doubt, more akin to the size of data that might be exposed in the industry rather than the size of the datasets that might be provided in a classroom setting.

1.2 Statistical Questions of Interest

The primary question of interest is how to predict Fraudulent Credit Card Transactions using Classification Algorithms. To achieve this however, due to the type of dataset, many statistical variables must be considered. **First is choosing the correct performance metric for evaluating accuracy.** By using a conventional accuracy metric, any accuracy score would be very misleading. This is due to the fact that the data is highly imbalanced, meaning, even if the model was to "always predict not fraud", it would be considered "99.83% accurate". **Second, is how to handle the problem of data imbalance.** While machine learning algorithms are typically robust for slight imbalance between classes, an imbalance to this degree makes it difficult for most algorithms. **Another consideration is best subset selection.** Due to the lack of context for each predictor, one must use statistical methods to select the best subset of the potential 30 predictors. The methods below will explore different techniques for doing this, such as L1 and L2 penalties applied to Logistic Regression, as well as exploring different subset selection methods. Additionally, there will be a discussion of the caveats of the various various selection methods.

2 Preprocessing

2.1 Identifying the Correct Performance Metric

By using the standard accuracy metric for Classification algorithms, the following confusion matrix reports an average accuracy of **99.93%**. A sample below is produced using a classic Logistic Regression algorithm with no data preprocessing and no tuned parameters.

```
[85293,     6]
[   52,    92]
```

While the total accuracy score seems high, the percentage of fraudulent transactions that were correctly identified is only **63.45%**, while the percentage of legitimate transactions that were correctly identified was **99.99%**. One can notice that, due to the imbalance of the classes, the

total accuracy is heavily misleading, as our goal is to be able to correctly identify fraud, and that was not done efficiently.

The best way to deal with this is instead of aggregating the accuracies, we look at them individually. We can use the Recall Score, which, in the example, is **63.45%**.The Formula for Recall is as follows:

Recall = True Positive/(True Positive + False Negative)

Additionally, it is good to keep in mind not just the % of Fraud that was predicted correctly, but also how good the prediction of the non-fraud was. This will be discussed in the next section.

2.2 Dealing with Class Imbalance

The next step is dealing with the class imbalance. As noted previously, running Logistic Regression on the raw, unsampled data outputs a Recall Score close to **63.45%**, which is unideal. This means that only **63.45%** of Fraud is being detected. Other algorithms such as LDA, and SVM also do not work well with high imbalance. While Tree Methods such as Random Forests are much more robust with class imbalance, a **0.17%** Fraud rate is daunting for any method.

To combat this, it is crucial that some form of Sampling method is applied. Upon different experimentation, we resided on comparing two different methods.

1. **Base Case:** Uses the same imbalance ratio to conduct the analysis
2. **Oversample and Undersample:** Undersamples the Larger Class, and Oversample the lower class to achieve a *ratio of 50:50*.

There are certain caveats that must be noted before conducting the sampling.

For the undersampling procedure, it is crucial to understand that there is potential for important information to be excluded. This is why the two performance metrics that was specified earlier is important. For the undersampling, it must be insured that the score that specifies what percent of non-fraud detected does not dip too far down. As a side note, given these caveats, it is clear that the sampling should be done without replacement.

The oversampling procedure, which is by nature a with replacement sampling technique, also has a potentially dangerous caveat. One must make sure that, when performing oversampling on a dataset, it must be done *after* splitting the data into Train and Test. This is true because the dataset is sampled with replacement. If the sampling is done *before* the split, then it is possible for duplicate samples to appear in both test and training sets, which would theoretically inflate the accuracy.

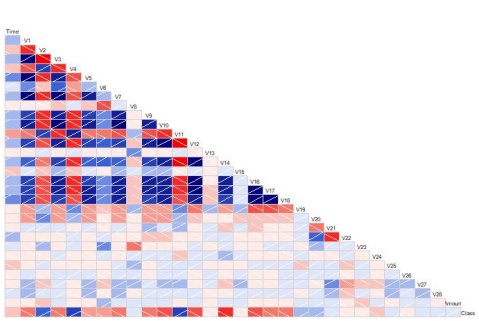
3 Exploratory Data Analysis and Visualization

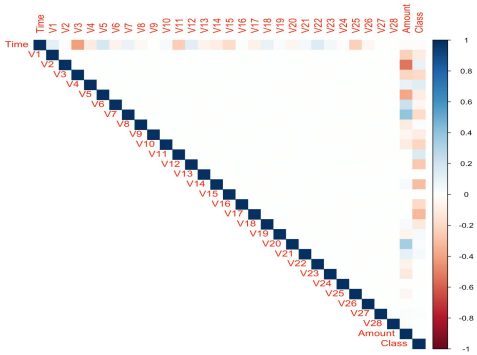
It is helpful to explore the data using various visualization techniques. On a side-note, while standardization of features may be useful before analyzing correlation between different predictors, this was not required in this specific scenario, as the data was run through PCA, which, by convention, is done after standardization.

Additionally, the estimated inferences from the plots will also be tested using more quantitative methods in later sections.

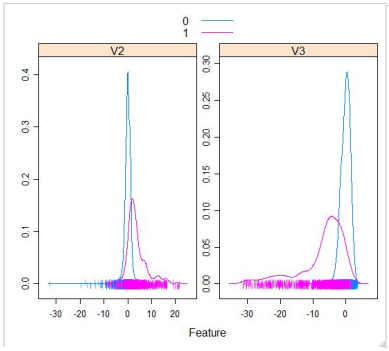
3.1 Distribution Plots of Class for each Potential Predictor

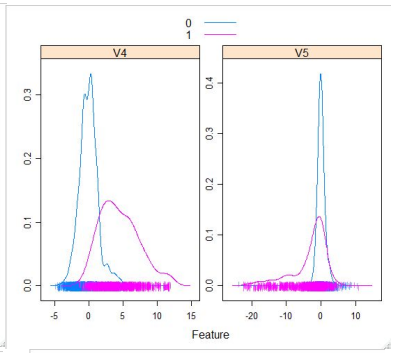
The following are correlation plots.
(Graph 3.1.1)

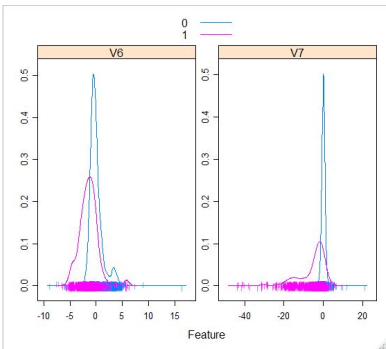


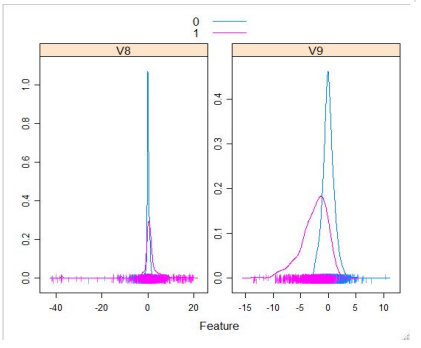


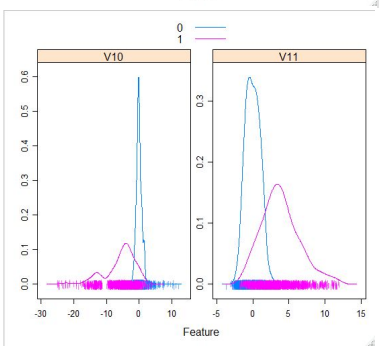
The following are the plots for each predictor mapped with Class. *(Graph 3.1.2)*

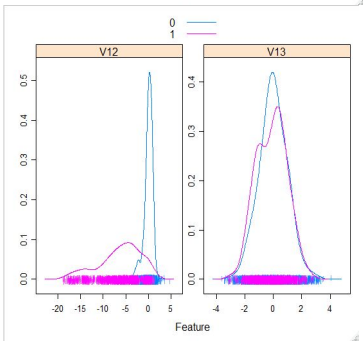


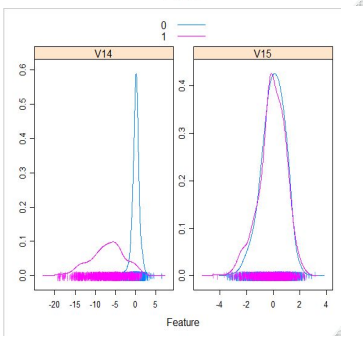












Note: The plot for the other predictors can be visualized by running the R code in the Appendix. Omitted here due to space issues.

Based on the descriptive analysis, we observe that:

1. The correlation plots below can give us an idea of which variables have strong or weak, positive or negative relationships with Class.(See in graph 3.1.1) Here, **V10, V12, V14, V16,** and **V19** have a fairly strong positive relationship with Class and **V2, V4**, and **V11** have a strong, negative relationship with Class. Using featurePlot, we can take a look at what variables in the dataset are less impactful to detecting fraud.
2. The featurePlots are used to analyze which variables are important in predicting whether or not transactions are fraudulent or non-fraudulent. The featurePlots consist of two different lines, the purple line represents Class 1 (fraudulent transaction) and the blue line represents Class 0 (non-fraudulent). (See in graph 3.1.2) Therefore the two classes above have a similar shape and frequency as one another. This therefore suggests that **V15, V20, V22, V24, V25**, and **V26** are less impactful on our target variable, Class.

Note: These estimations will be re-evaluated in the model building process

4 Modeling and Prediction

To evaluate model accuracy score with the basis of context, we can set a **base case** for our accuracy (Using the performance metric we established earlier). We refer to the model specified earlier to establish our base case:

Base Model: *Accuracy Score ~63%*

- All Predictors
- Raw Dataset (no sampling methods)
- Logistic Regression (no penalty)

The following modeling methods are built with the conditions and constraints discussed in the previous sections. It will be in our best interest to attempt to achieve an accuracy stronger than **63%**. To handle the preprocessing efficiently, we wrote various functions to simplify the process. While the preprocessing function can be observed in the Appendix, the following are the procedure that occurs each time the code is run:

1. Split Train and Test set, 70:30 ratio (or else if specified in the parameter)
2. Applies Sampling Method, either a 1 or 2, details specified in earlier sections
3. Outputs the new Data, split between Train and Test

With modeling, the above mentioned function will be wrapped by the functions used by the Algorithm.

4.1 Logistic Regression (with Random Forests)

Logistic Regression is a powerful classification algorithm. It is typically faster than most other classification algorithms, although Tree Methods such as Random Forests seem to often win the race. It also, theoretically speaking, performs better than algorithms such as LDA when the conditions are not met for LDA. As an example, referring to *Introduction to Statistical Learning*, LDA performs better than Logistic Regression when classes are well separated, as well as if the distributions of the classes look normal. As these assumptions feel violated for the most part, Logistic Regression is a clear first consideration. Other algorithms, such as SVM (Support Vector Machines) is infamous for their atrocious computational speed, thus, would not be an initial consideration as our dataset is fairly large.

Alongside Logistic Regression, Random Forests were used as a method for extracting the most significant predictors. Random Forest Feature Extraction is one of the methods we used for modifying the Dataset.

Note: To evaluate the accuracy of a model type, ***each model type was run numerous times*** *and thus the reported accuracy scores are the average of each model type.*

Note 2: From here on out, each model will specify the accuracy of the Fraud, and then followed by the accuracy of the non-Fraud, as discussed earlier.

4.1.1 Applying Logistic Regression (No Penalty) on Balanced Data

The following are Logistic Regression Models ran with no penalty and on both the balanced dataset and base model for comparison.

Model 4.1.1 (Base): *Accuracy Score ~58%, 99.98%*

- All Predictors
- Raw Dataset (no sampling methods)
- Logistic Regression (no penalty)

Model 4.1.1 (Balanced): *Accuracy Score ~90.32%, 97.73%*

- All Predictors
- Balanced Dataset using Oversampling and Undersampling
- Logistic Regression (no penalty)

4.1.2 Applying Logistic Regression (L1 and L2 Penalty) on Balanced Data

The following are Logistic Regression Models ran with L1 and L2 regularization. L1 and L2 are popular methods to - in a sense - , choosing the best predictors. How it actually works is that the least significant predictors are minimized or eliminated, depending on which penalty is used.

Model 4.1.2 (Balanced, L1 Penalty): *Accuracy Score ~90.24%, 97.71%*

- All Predictors
- Balanced Dataset using Oversampling and Undersampling
- Logistic Regression (L1 Penalty)

Model 4.1.2 (Balanced, L2 Penalty): *Accuracy Score ~92.10%, 97.72%*

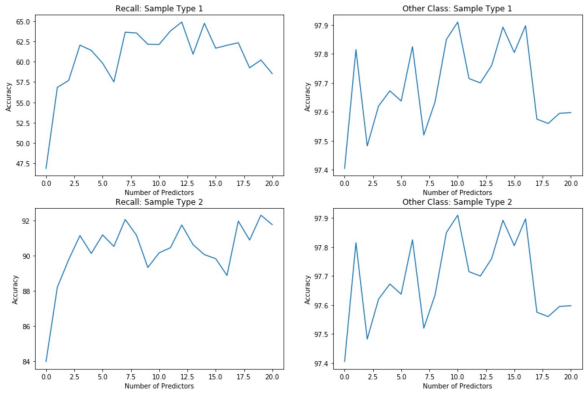
- All Predictors
- Balanced Dataset using Oversampling and Undersampling
- Logistic Regression (L2 Penalty)

*Note: Logistic Regression with L2 Penalty ran **significantly** faster than L1 or base*

4.1.3 Logistic Regression, using Random Forests for Feature Selection on Balanced Data

A popular method for feature selection is using Random Forests to extract the most significant features. According to *Introduction to Statistical Learning*, this is done through comparing tree node impurity along with the number of samples that reach the particular node.

Note: The feature importance can be seen in the Appendix in the Python section



Upon finding the order of feature importance, we can use a loop to find the best number of predictors to include, using Recall Accuracy and use a visual criteria such as the Elbow Method like the KNN, because parsimony is also a factor in this particular model selection. This is visualized on the left.

The right graphs are less significant. Carefully notice that the scales are different. Using the elbow criterion for the left graphs, we can infer that using the **top 3** predictors may create a good model.

Model 4.1.3 (Balanced, RF Feature Selection): *Accuracy Score ~91.14%, 97.71%*

4.2 Other Model Considerations

4.2.1 Other Algorithms

Other algorithms were considered and experimented, and some can be observed in the appendix. However, the best model we were able to produce was Logistic Regression, with extensive parameter tuning.

4.2.2 Caveats with Best Subset Selection

In the textbook, *Introduction to Statistical Learning*, subset selection using forward and backward stepwise selection is emphasized and encouraged. However, multiple sources and forums online seem to not prefer this technique, especially criticizing the criterion used, AIC, AICc, or BIC. Thus, we decided to step away from this approach. In addition, there was an instance where AIC/BIC criterion could have been utilized, however we chose a more computationally expensive method using cross-validation to ensure optimal accuracy.

5 Discussion

The results of our study provide not only an accurate method of predicting credit card detection, but also the general means to do so, using various functions that was written with the group. There is a deep and careful exploration of data preprocessing techniques, and each method that is explored is heavily insured with both benefits and major caveats.

In our findings, we were able to raise our base 63% accuracy to around 92%, using Logistic Regression with L2 Penalty using an Undersampled/Oversampled dataset. This was a challenging project as we encountered many problems not covered in the course, such as dealing with many predictors, high class imbalance, and redefining performance metrics.

Given more time, resources, and knowledge of Machine Learning algorithms, a future consideration is to revisit this project using Isolation Forests and Gradient Boosting. Upon light exposure, it seems that Isolation Forests specialize in anomaly and fraud detection, and Gradient Boosting methods such as XGBoost seems to be at the forefront of Machine Learning, alongside Deep Learning.

6. Appendix

Note: All material from this point on is for the Appendix.

Sources used:

https://www.rdocumentation.org/packages/Seurat/versions/3.1.1/topics/FeaturePlot
https://cran.r-project.org/web/packages/unbalanced/unbalanced.pdf
https://data-flair.training/blogs/data-science-machine-learning-project-credit-card-fraud-detection/

https://stackoverflow.com/questions/14463277/how-to-disable-python-warnings
https://pandas.pydata.org/pandas-docs/stable/user_guide/merging.html
https://medium.com/datadriveninvestor/rethinking-the-right-metrics-for-fraud-detection-4edfb629c423
https://stats.stackexchange.com/questions/20836/algorithms-for-automatic-model-selection/20856#20856
https://matplotlib.org/3.1.0/api/_as_gen/matplotlib.pyplot.subplots.html
https://www.geeksforgeeks.org/confusion-matrix-machine-learning/
http://faculty.marshall.usc.edu/gareth-james/ISL/
https://stackoverflow.com/questions/25427650/sklearn-logisticregression-without-regularization
https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3