Analyzing Supervised Learning Methods for Credit Card Fraud Detection

Jennifer Gao
D. W. Daniel High School
MehtA+Tutoring

Karen Situ
University Hill Secondary School
MehtA+Tutoring

Grace Tian
Yorktown High School
MehtA+Tutoring

July 23, 2020

Abstract

Credit card fraud has been a growing issue both in the Unites State

1 To-do (other than proofreading), will delete later

- Karen finish ur svm and stuff, also do random forest for back up
- some images are blurry and fonts in images are small (I could probably convert them to some latex-ed version so zooming in works)
- abstract
- i heard we're lacking content? aka we should write more
  - but one way to make it look like we have more content is to make the subsubsections unnumbered
  - is it a good idea to explain how each model works?
- combined model stuff
- conclusion
- future work
- division of labor
- acknowledgements?
- there's this one random table on the last page but im gonna leave it there for now in case we need to copy paste later

2 Introduction

Every year, billions of dollars are lost worldwide due to credit card fraud. In the U.S. alone, 9.47 billion dollars were lost in 2018, and this amount is projected to increase in the coming years [1]. By analyzing the patterns of current credit card fraud data, future fraudulent credit card transactions could be predicted in advance and stopped.

- Unsupervised: A Comparison Study of Credit Card Fraud Detection: Supervised versus Unsupervised
- Decision tree: Predictive Modelling For Credit Card Fraud Detection Using Data Analytics
- KNN: Analysis on credit card fraud identification techniques based on KNN and outlier detection
- Forest: Credit Card Fraud Detection Using Random Forest Algorithm
- Comparison: Analysis of Credit Card Fraud Detection Techniques: based on Certain Design Criteria

Past studies have used supervised learning with decision trees (DT), KNN, and random forest (RT) individually and have compared their effectiveness. However, there has not been evaluation of a combined DT-KNN-RF model. By combining the three models using various methods, we were able to obtain higher precision and recall rates than any of the three models alone yielded.

3 Methodology

3.1 Dataset

The "Credit Card Fraud Detection" dataset from Kaggle consists of 492 fraudulent and 284807 nonfraudulent transactions, collected within a span of around 2 days [2]. Note that this dataset is heavily imbalanced, with less than 0.2% of the transactions being fraudulent. The 30 features include time in seconds after the first transaction in the dataset, transaction amount, and 28 other anonymized features representing sensitive data after principal component analysis (Figure 1).

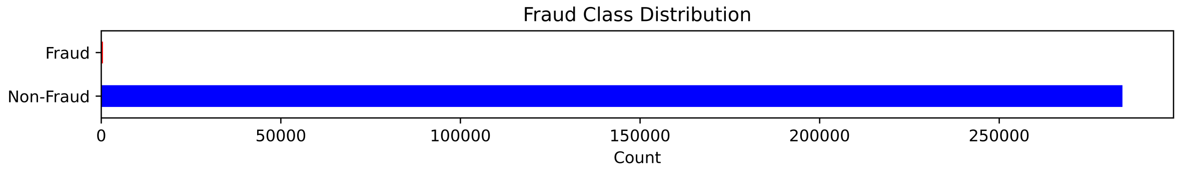

Figure 1: Imbalance in fraud class distribution.

Preprocessing In order to deal with the heavily imbalanced data classes, we use synthetic minority over-sampling technique (SMOTE), which generates synthetic data in the minority class to create a class balance for unbiased training [3]. We first split the imbalanced data using a 67:33 train:test split. This allows us to achieve the final desired 80:20 train:test split after generating more fraudulent data to train on using SMOTE (Figure 2).

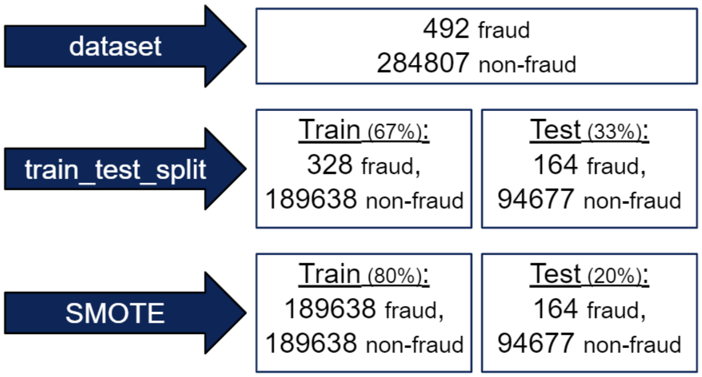

Figure 2: Use of SMOTE to balance testing data.

3.2 Models

We developed three separate supervised classification models: decision tree, k-nearest neighbors, and random forest. We then combine the three models to create a create a more robust model with high sensitivity to fraud.

Decision Tree

We built a decision tree classifier with a 600:1 class weighting to reflect to approximate ratio of nonfraud:fraud. After conducting experiments on the max_depth parameter, we found that a maximum tree depth of 6 gave the highest recall without compromising precision. Although some higher max_depth yielded greater recall values, these increases were minor compared to the large decreases in precision, as the model begins to overfit after a maximum tree depth of 6 (Table 1).

KNN Classifier

We built a KNN classifier using a distance weight to put a larger emphasis on closer points. After conducting experiments varying the number of neighbors, we found that using four nearest neighbors optimized recall rate, so we used four nearest neighbors to classify our test data.

Table 1: Evaluating decision tree at different values of max_depth.

max_depth	Precision (%)	Recall (%)
1	76.8	64.6
2	82.3	70.7
3	83.7	72.0
4	88.3	78.0
5	89.7	79.3
6	**91.0**	**80.5**
7	87.2	79.3
8	85.2	80.5
9	84.4	82.3
10	81.7	81.7

Random Forest

A random forest algorithm has its advantage of not overfitting, which comes useful especially when training on imbalanced data. Random forests work by taking the most frequent results given by a set of decision trees. After testing, we have decided to implement the random forest classifier using 800 trees.

DT-KNN-RF Combined Model

hihihi

4 Results

4.1 Decision Tree

Our final decision tree had a recall of 80.5%, which means it successfully detected 80.5% of true frauds. It also had an accuracy of 99.95% which is due to the imbalanced test set. The decision tree's precision was 91.0%, so 91.0% of predicted frauds were actually frauds (Figure 3).

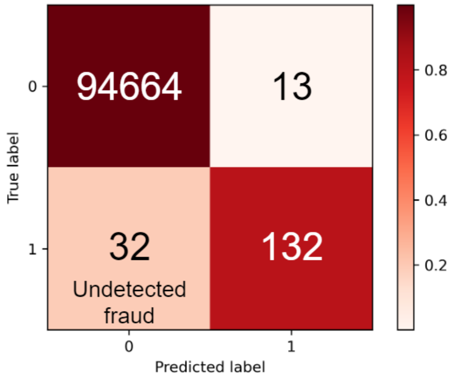

Figure 3: Confusion matrix for decision tree.

Due to the high interpretability of decision tree models, we analyzed the feature importance in the decision tree to determine which features were most influential in the classification process. From our analysis, we see that feature pc17 is the most important, which is confirmed since feature pc17 decides the first node in the decision tree (Figure 4). Although the pc17 feature is a result of anonymizing original sensitive features under principal component analysis, future research can look into uncovering the original sensitive features which are correlated to pc17. This could lead to more efficient methods of credit card fraud detection which are based on observable patterns in fraudulent transactions.

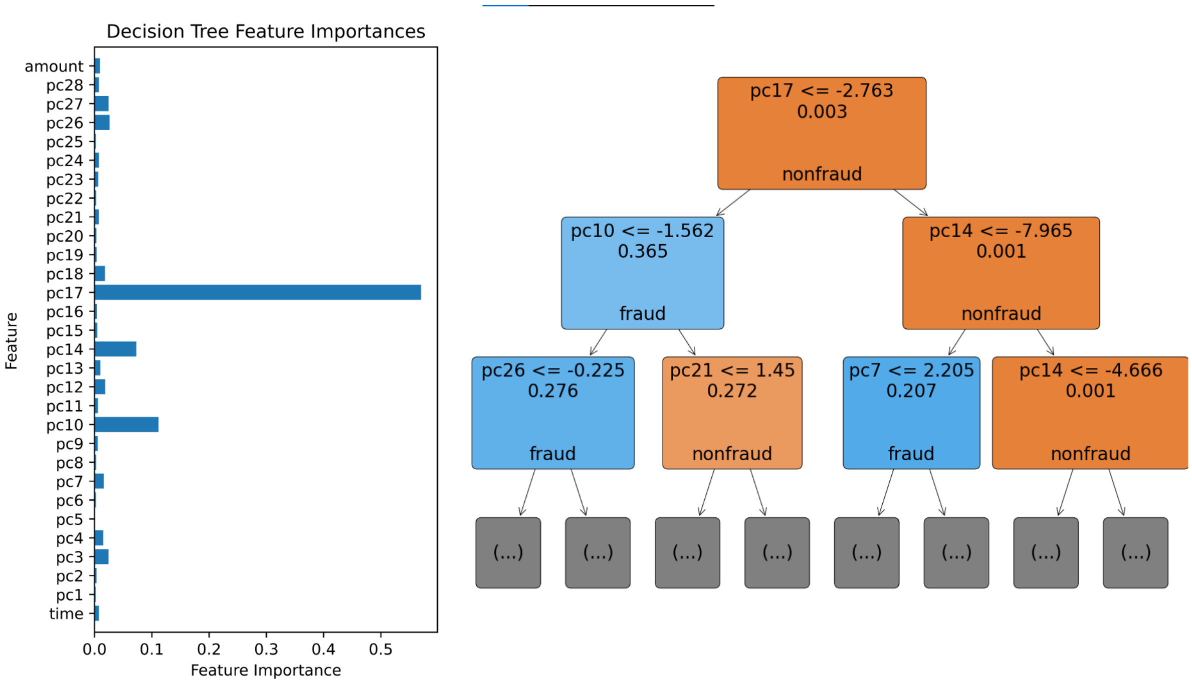

Figure 4: A graph of feature importances in the decision tree (left). A visualization of the decision tree at a depth of 2, where orange represents nonfraud and blue represents fraud (right).

4.2 KNN Classifier

Our KNN model had a recall rate of 56.1%,meaning it successfully detected 84.1% of true frauds (Figure 5). It also had a 95.74% accuracy rate. While this model has a lower recall and precision rate than the other two models, when combined with either, it yields a higher recall rate than any of the models alone.

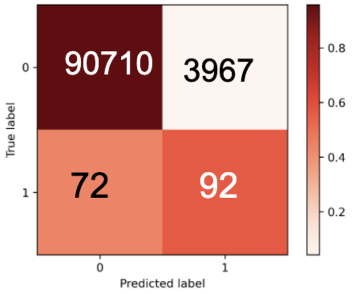

Figure 5: Confusion matrix for KNN classifier.

4.3 Random Forest

qwertyuiop

4.4 Combined Models

5 Conclusion and Future Work

Please add conclusion here.
This creates a new line.

Table 2: Comparison of

Part		
Name	Description	Size (μm)
Dendrite	Input terminal	$\sim$100
Axon	Output terminal	$\sim$10
Soma	Cell body	up to 10^6

The command above this line created some space.

6 Division of Labor

We divided the work as follows:

- Item 1
- Item 2
- Item 3

7 Acknowledgements

We would like to acknowledge Haripriya Mehta, Bhagirath Mehta, Marwa AlAlawi, and Andrea Jaba for their help in teaching and advising us throughout this project.

References

[1] The Nilson Report. Payment card fraud losses reach $27.85 billion, Nov 2019.

[2] Machine Learning Group ULB. Credit card fraud detection, Mar 2018.

[3] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.