KDDM Project 2: Fraud Detection

Group Model Summary

Zero Credibility

April 10th, 2022

ID 5059

1 Model Description

1.1 Modelling Strategy

To tackle this important fraud detection problem, the Zero Credibility team divided and conquered. We split into individual units approaching this problem from multiple angles: Logistic Regression, Bagging, Trees, and Boosting. Each member independently developed their own model and submitted it to test the efficacy on an unseen dataset to obtain the greatest predictive accuracy. Logistic Regression had greatest interpretability but lacked accuracy, Bagging and Tree models were easiest to visualize feature relationships, but lacked flexibility, and Boosting dominated in prediction accuracy. Once the final data were aggregated it became obvious that the most efficient and effective model was one using a Boosting algorithm.

1.2 Final Model

Our final model with the greatest predictive power was developed using a method called “Extreme Gradient Boosting”. The boosting algorithm was most effective in detecting credit card fraud because it fits many models to the data and “learns slowly”. From each model it boosts the weights of observations predicted wrong by the previous learner, re-samples based on the new weights and fits a new learner. It was internally validated using the “ROC-AUC” method. We validate using this score in a binary classification problem because it quantifies the outcomes with a score between 0 and 1. Intuitively, a score closer to 1 will illustrate a model that has adequately identified differences between the classes, 0.5 indicates a random guess, and 0 is a complete crossover.

Validating the data on an unseen validation set, our model yielded a mean score of 0.95 which in turn yielded an impressive 93% accurate fraud detection rate on unseen data. To achieve this accuracy, the model is using a technique called cross validation, where it splits the training data in 6 parts, fits the model on 5 parts and uses the 6th part to assess the accuracy. Every time the model is being fit, the model will also predict the probabilities for the test data and at the end, the 6 sets of probabilities will be averaged. From this model we determined that the most important features were "V258, V257, V188 and V70", and we would suggest the data provider should ensure this data is always being collected for future fraud detections. Additionally, the client should enhance the data set with credit card identifiers (where available) that can help analyze transactions. While Zero Credibility has done work to group transactions together, this solution is not bulletproof and it would help achieve an even better accuracy than the current one.

2 Business Case

The necessity for fraud detection cannot be overstated. In research conducted by the University of Portsmouth's Centre for Counter Fraud Studies, it was determined: "fraud is costing the global economy £3.89 trillion, with losses rising by 56% in the past decade" [3]. As statisticians and computer scientists, we have the unique opportunity to drive change in the global marketplace and aid in the reduction of hemorrhaging. We have built a robust, efficient, accurate, scalable model to solve this major need. Take the example, Bank of Scotland who deals with over 1 billion transactions a year, and a mean loss from fraud being £10, increasing the detection rate by just 0.1% can yield an annual saving of £1,000,000 [1].

Leave the detection to Zero Credibility's Fraud Detection Algorithms. Our abilities to identify, track, and notify you about potential fraudulent transactions, faster and more accurately than any other group in ID5059 is what sets us apart from the rest! Credit Card fraud is on the rise and with the current economic climate, every dollar matters. In the past year we have seen unprecedented rises in credit card fraud (48%) [2] and you need us. Our machine learning models leverage new levels of artificial intelligence that can detect patterns in seemingly normal data. The ability to synthesize data that the average person would see as nothing.

3 References

1. Bolton, Richard J., and David J. Hand. "Statistical Fraud Detection: A Review." Statistical Science, vol. 17, no. 3, 2002, https://doi.org/10.1214/ss/1042727940.

2. Daly, Lyle. "Identity Theft and Credit Card Fraud Statistics for 2021: The Ascent." The Motley Fool, The Ascent by The Motley Fool, 27 Oct. 2021, https://www.fool.com/the-ascent/research/identity-theft-credit-card-fraud-statistics/.

3. "The Financial Cost of Fraud 2019." Crowe UK, https://www.crowe.com/uk/insights/financial-cost-of-fraud-2019