**Executive Summary**

**D214 Capstone**

**Jun 02, 2022**

**Statement of problem and hypothesis**

Fraud is a continuous problem that affects more than 2.8 million consumers every year (FTC, 2022). The use of statistical tools and data analysis has been helping to ease some of these difficulties, but problems persist. In this analysis, we are looking at synthetic consumer fraud in Europe in 2021 and seeing if using Naïve Bayes techniques and using smaller sample sets calculated via student’s t-test. This technique will help improve the detection of fraud and be comparable to using a smaller dataset to help save time, processing power, and storage space. The simple question that this analysis aims to answer is “Can we detect fraudulent charges in a transactional dataset?”. After narrowing that question down, we could also ask “How well does Naïve Bayes work at predicting fraudulent charges in a transactional dataset?” The hypothesis is that yes, we will be able to predict fraudulent charges using a Naïve Bayes analysis. The answer to the second question is yes, we will be able to select an adequate sample size to accurately predict fraudulent charges in the dataset.

**Summary of Data-analysis process**

The full purpose of this analysis is to see how accurate a smaller, statistically significant, sample size can be when compared to an entire big dataset. We will do a little visualization of the dataset and then we will begin to apply naive bayes theorem to these datasets to try and predict fraud. Naive bayes works by taking previous results and using it as a predictor for the next events. Naive bayes works well if all the components are independent of each other. We decided to clean the data up by deleting duplicates and then created a small visualization to see what kind of data we are working with. We then decide to drop a couple columns that weren’t important and then ran a correlation test between all the remaining variables of type, amount, oldbalanceOrg, newbalanceOrig, oldbalanceDest, newbalanceDest, isFlaggedFraud, and size and to compare those to the isFraud column to see if any relationships could be determined. After not really finding many we still decided to run a naïve bayes analysis to see if it could predict the outcome of fraud and it was overfitted, so the model was dropped. Our next test was to create an adequate sample size using Cochran’s formula which came out to a set of 16590 samples. We extracted that, then created models based on the sample set, and the original dataset without the samples.

**Outline of Findings**

After running a naïve bayes prediction analysis, we created a model of the original data of accuracy 99.2% and using the sample dataset, we got an accuracy of 53.8%. It appears that at a cost of time and space, we can get higher accuracy, but sacrificing that, our analysis is little better than a coin flip.

**Limitations**

Using the technique of naïve bayes for a predictive analysis does have a limitation of the assumption of independent predictors. One other limitation was the sample size that was created. Using 16590 samples is what was calculated using a Cochran’s formula estimation. There are other options we could use for sample selection but if the sample size is over 16590, we should have at least 99% confidence with a 1% margin of error.

**Proposed Action**

The best course of action would be to run another analysis in order to get a better understanding of the data. If we were to perform a logistic regression and a decision tree analysis, we could have some back up results that could help guide our executive decision making. The more tests that are run, the more accurate our results can be, and we can make even better decisions that would affect the stakeholders.

**Expected benefits**

For this analysis we asked ourselves if we could accurately predict if a transaction was fraudulent or not by applying a Naive Bayes analysis to the dataset and train a sci-kit machine learning tool to detect that. We also decided to try seeing if we could speed up analysis time by pulling out a smaller sample set to apply our analysis to and see if it'd perform similarly to help save computation time and space. After performing our analysis that is demonstrated above, we were able to create a system that uses the bulk dataset to create a 99.25% accurate predictor of whether a transaction is fraudulent or not. This came at a cost to the size and time that it took to analyze almost 6.2 million records in order to create our model. We also created a system using a statistically selected sample set to see if we could speed up our time while also saving space. This sample model came back at 53.83% which isn't fantastic, but it was able to predict our fraudulent charges correctly for all our test cases.

**Panopto Video:**

**Resources**

<https://www.ftc.gov/news-events/news/press-releases/2022/02/new-data-shows-ftc-received-28-million-fraud-reports-consumers-2021-0> , FTC, 2022

<https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc> , Saishruthi Swaminathan, 2018

<https://scikit-learn.org/stable/modules/tree.html> , Sci-Kit, 2022

<https://www.kaggle.com/datasets/rupakroy/online-payments-fraud-detection-dataset> , Kaggle, 2022

<https://analyticsindiamag.com/why-is-overfitting-so-demonized/#:~:text=%E2%80%9CDue%20to%20overfitting%2C%20a%20model,unseen%20data%20in%20real%20situations> , Shraddha Goled, 2021

<https://www.tarleton.edu/academicassessment/documents/samplesize.pdf> , Glenn D. Israel, 2003

<https://www.tdistributiontable.com/> , T-Table, 2021