Approach Document for Nemesis Consultants LLP – Fraud Detection

A brief on the approach, which you have used to solve the problem.

I've used Python in Jupyter Notebook and spark as well to make a reusable and scalable code that would help our Data science team, and in turn, help client – to get a clean data source to be used for model input and get predicted Fraudulent Transactions

The code generally helps in data cleaning, addressing the important features and predicting the accurate results.

What data-pre-processing / data cleaning ideas really worked? How did you discover them?

Since this was a simple but dirty data, the path to the right solution was muddy. Overall, what really helped was to start thinking why the data is incorrect, and what all could work to make it better.

- For all records, I filtered the duplicate Records
- Secondly, there were a lot of Null values and missing string values in columns.
- Third, I made sure that the data isn't having any missing Values
- For Activity, we loaded a sklearn.preprocessing technique i.e. Label Encoder to encode our categorical feature as well to get more accurate predictions.
- And then split our datasets into test and train to get performance measure
- Finally, I've used scikit-learn package to get binary classified result of isFruad column

After all the data cleaning/imputations, I broke down the dataset to calculate all input features separately and, in the end, I made two Predictive models.

1. Logistic Regression
2. Bernoulli's Naïve Bayes

This not only makes the code reproducible, but also makes sure it's easier to predict the isFruad with more Accurate results

Which tools did you use to solve the problem?

I used Python with sklearn, Pandas, Numpy, matplotlib, seaborn libraries.