Q.1 Credit Card Fraud Detection.

Q.2 The topic I choose to apply the data science methodology is "Credit Cards"

Q.3.

1. According to the World Payments Report, in 2016 total non-cash transactions increased by 10.1% from 2015 for a total of 482.6 billion transaction. And in future years there will be a steady growth of non-cash transactions. With that fraudulent transactions are on the rise as well. We have a very high amount of money lost from credit card fraud.
2. So how can we detect a credit card fraud?

Q.4

1. Analytic Approach:

For the detection of fraud in credit card we use Predictive approach.

2. Data Requirements:

For Predictive approach we need data so that we can apply algorithms. so we need datasets contains transactions made by credit cards.

3. Data Collection:

I gathered my data from a Kaggle dataset which contained 285,000 rows of data and 31 columns. Out of all the columns, the only ones that made the most sense were Amount, Time and Class (fraud or not fraud). The data itself is short in terms of time (it’s only 2 days long), and these transactions were made by European cardholders.

4. Data Understanding and Preparation:

Lets just explore the important features like Time, Amount, and Class.

TIME:

Most purchases are made during the daylight hours, and as people get out of work/school and head home, purchasing dwindles down until the next day.

AMOUNT:

Vast majority of transactions are very low.Most daily transactions aren’t extremely expensive (most are <$50), but it’s likely where most fraudulent transactions are occurring as well.

CLASS:

There are only 492 fraudulent transactions. That’s only 0.173% of all of the transactions in this dataset.

and also search for the features that’s showing high correlation.

STANDARDSCALER:

StandardScaler transforms the data to where there is a mean of 0 and a standard deviation of 1, thus standardizing the data into a normal distribution. Here i apply it to Time and amount feature.

5. Modeling and Evaluation:

Here I decided to a 60% train/20% validation/20% test split dataset. And after that apply Logistic Regression, Random Forests and Naive Bayes the Accuracy are:

Logistic Regression: 0.919, Random Forests: 0.924, and Naive Bayes: 0.874

Logistic Regression and the Random Forest models are very promising for our dataset. Each model has a high true positive rate and a low false positive rate. Both models are still promising and give good results in the end.