Assignment, IBM Data Science Methodology (Coursera)

Credit Card Fraud

1. Which topic did you choose to apply the data science methodology to? (2 marks)

The topic that I have selected to apply data science methodology to is Credit Card Fraud. I chose this topic as credit card fraud is a significant concern for both financial institutions and consumers. Detecting fraudulent transactions is crucial to prevent financial losses and maintain the trust of customer.

2. Next, you will play the role of the client and the data scientist.

Using the topic that you selected, complete the Business Understanding stage by coming up with a problem that you would like to solve and phrasing it in the form of a question that you will use data to answer. (3 marks)

You are required to:

1. Describe the problem, related to the topic you selected.

2. Phrase the problem as a question to be answered using data.

For example, using the food recipes use case discussed in the labs, the question that w defined was, "Can we automatically determine the cuisine of a given dish based on its ingredients?".

Credit card fraud is a significant concern for both financial institutions and consumers. Detecting fraudulent transactions is crucial to prevent financial losses and maintain the trust of customer.

Problem Definition: To predict whether a credit card transaction is fraudulent or legitimate based on historical transaction data.

Question: Is it possible to automatically classify credit card transactions based on various features and patterns extracted from the transaction data?

3. Briefly explain how you would complete each of the following stages for the problem that you described in the Business Understanding stage, so that you are ultimately able to answer the question that you came up with. (5 marks):

1. Analytic Approach
2. Data Requirements
3. Data Collection
4. Data Understanding and Preparation
5. Modeling and Evaluation

You can always refer to the labs as a reference with describing how you would complete each stage for your problem.

1. Analytic Approach: I will use predictive approach (Predictive Analysis –Classification). As the problem requires a yes/no answer (fraudulent/ legitimate) I will use a classification model. I will build models to predict whether the transaction is likely to be fraudulent or not.

2.Data Requirements: To build a model for automatically classifying credit card transactions, I might need the following types of data:

- Historical Transaction Data (a record of past credit card transaction): Such as the transaction amount, transaction frequency, merchant name, transaction location, timestamp, etc
- Cardholder Data: Such as demographic data, historical spending patterns, etc.
- Fraud Label Data: A labeled dataset indicating which transactions are fraudulent and which are legitimate.

3.Data Collection: Data can be obtained from several resources such as financial institution’s transaction records, public datasets (i.e. datasets on Kaggle website), data providers (i.e. companies specializing in providing data for various purposes, including fraud detection), etc.

in order to make sure I have relevant data, I will do some data exploration (i.e. descriptive statistics & visualization). This will give me initial insight, help me assess content and quality, and identify gaps. This is to ensure that I have useful data for my model.

4.Data Understanding and Preparation:

Data Understanding: I will look at quality of data (e.g. missing values, invalid/misleading values), and completeness of data. I will also do descriptive analysis (e.g. histogram, univariate analysis, pairwise correlation) to see what data preparation will be needed.

Data Preparation: I will clean and preprocess the data to ensure its quality and readiness for analysis. These steps may include, handling missing values, duplicates, invalid entries. I might need to some data transformation (e.g. normalize or standardize numerical features to ensure consistent scaling), and feature engineering (i.e. create new features)

5. Modeling and Evaluation

Modeling: I will build a predictive model. I will train machine learning models using features such as historical data. The models learn to differentiate between legitimate and fraudulent transactions based on patterns inherent in the data. As the model encounters new transactions, it assesses these features to predict whether the transaction is likely to be fraudulent or not.

Evaluation: Then, I will assess the quality of the trained models using evaluation metrics such as accuracy (i.e. proportion of correctly classified instances), precision, recall, and F1-score. If the model's performance is not satisfactory, I will fine-tune hyperparameters and try different algorithms. Finally, I will use ROC curve to determine the optimal classification model.