Data Methodology Case Study

Which topic did you choose to apply the data science methodology to?

ANS: I choose the credit card as the topic where we can use the information of credit cards to avoid the credit card fraud.

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.

You are required to:

Describe the problem, related to the topic you selected. 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 we defined was, "Can we automatically determine the cuisine of a given dish based on its ingredients?"

Answer: It is important that the credits companies are able to identify fraudulent credit card transactions. The credit card and online net banking fraud is an international problem in banking domain. The global fraud accounted for loss is increasing day by das fraudster are developing new techniques to alter the normal working behavior of the credit card. One of the main challenges is how to improve fraud detection accuracy with the growing number of transactions done by users per second.

Question: Base on the credit card fraud, is it possible to propose a solution by building analytical framework model for fraud detection.

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):

It is important that the credits companies are able to identify fraudulent credit card transactions. The credit card and online net banking fraud is an international problem in banking domain. The global fraud accounted for loss is increasing day by das fraudster are developing new techniques to alter the normal working behaviour of the credit card. One of the main challenges is how to improve fraud detection accuracy with the growing number of transactions done by users per second.

Question: Base on the credit card fraud, is it possible to propose a solution by building an analytical framework model for fraud detection?

Analytic Approach

Data Requirements

Data Collection

Data Understanding and Preparation

Modeling and Evaluation

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

Answer:

1.Analytic approach

There are many data science algorithms that can be applied to data. In this case, the model I want to build is used to check if incoming transaction is legal transaction or not. The predictive model such as logistic regression and decision tree can be implemented for fraud detection.

2. Data requirements

In the case, the first task was to define the data required for the classification approach of the selected logistic regression model. It involved selecting a suitable data from different sources.

The datasets should contain the transactions made by credit cards in specific period, for example in last two months, by cardholders. It is necessary to define the content, format of the data sets which will be used to classify the logistic regression model.

3. Data Collection

In the data collection process, it is necessary to review if the data requirements are successfully achieved and make a decision whether more or less data is required for the collection.

The collected dataset could be in an unstructured form so it would be necessary to define the fields including the time, amount as well as class. Time shows the time gap between the first transaction and the following one. Amount is the amount of money transacted. Class represents the classification of the purchased items. It would be necessary to remove that unwanted information. These data are recognized as attributes of this problem.

4. Data Understanding

Afterwards, the data scientist should be aware of the its the collected data representative of the problem they are trying to solve. To gain a better understand of the collected data, techniques like descriptive statistics can be implemented in the data to show the variables of the predictive model. The graphs such as the amount, time and class of the transaction can plotted. Then the variables such as mean, median, minimum, maximum in each graph can be used to understand the distribution. This is done to get a graphic representation of the dataset which can be used to verify that there are no missing values in the dataset and also to ensure the machine learning algorisms can process the dataset smoothly.

5. Modeling and Evaluation

By using various developed predictive model with prepared training data, the evaluation of the model needs to perform to check if the model can be used to predict credit card fraud. A comparison between a set of predicted results by various machine learning algorisms with known results can be applied by plots and tables to ensure the accuracy of the analytical model. By analyzing the predicted dataset used for model refinement.