MIS 637 Midterm
Data Analytics and Machine Learning

October 17, 2019
School of Business
Stevens Institute of Technology

Professor M. Daneshmand

Student Name: SRUJAN JADHAV

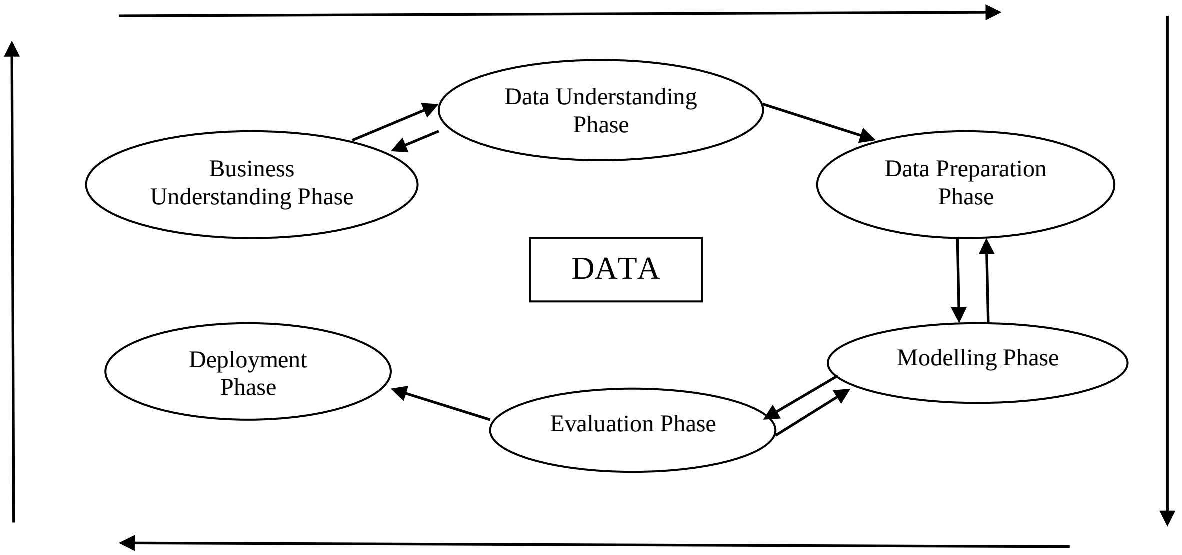
A mortgage company likes to be able to decide on 3 interest rates for new loan applicants as follows: an interest rate of 5% for “high risk” applicants, 3% for “average risk” applicants, and 2% for “low risk” applicants. You are in charge of this project. Provide a comprehensive end-to-end plan for this project. Include all the necessary steps from the beginning to the end. Make any necessary assumptions and define notations. Give a comprehensive description of the algorithm(s) as well as the related formulas you will use for this project (this is the fundamental part of your role in this project). Provide a detail description of the algorithm and how does it work. Please put your answer in the format of Step 1, Step 2 …

Answer:

Given data: 1. Interest rate of 5% for “high risk” applicants
2. Interest rate of 3% for “average risk” applicants
3. Interest rate of 2% for “low risk” applicants

An end-to-end plan for the project is created using CRISP-DM model. In this project, CRISP-DM is used to decide which applicant gets which loan plan. Furthermore, CRISP-DM model can be used to carry out promotional events about the loan.

CRISP-DM stands for Cross Industry Standard Process for Data Mining. The CRISP-DM model consists of 6 important steps.



MIS 637 Midterm
Data Analytics and Machine Learning

October 17, 2019
School of Business
Stevens Institute of Technology

Professor M. Daneshmand

Student Name: SRUJAN JADHAV

They are:
Step 1. Business Understanding Phase
Step 2. Data Understanding Phase
Step 3. Data Preparation Phase
Step 4. Modelling Phase
Step 5. Evaluation Phase
Step 6. Deployment Phase

STEP 1: BUSINESS UNDERSTANDING PHASE

In Business Understanding Phase, the project objectives and requirements are put forward in terms of business unit. These goals are then translated into fata mining problem definition. Then a strategy is created to achieve these objectives.
In this project, the main problem is to figure out which loan plan is best for the customers.
Here the given data is the requirement: 1. *5% for "high risk" applicants*
2. *3% for "average risk" applicants*
3. *2% for "low risk" applicants*

STEP 2: DATA UNDERSTANDING PHASE

In Data Understanding phase, the actual data is collected. Then EDA (Exploratory Data Analysis) is performed on collected data and insights are found out. The quality of the data is checked.
In this project, the data is visualized into statistical data, various insights like relation between the applicant and the loan plan is found out, credit history of the applicant will be checked to decide the risk factor of the applicant. Then data is cleaned or missing value treatment is carried out. The data is made ready for further analysis.

STEP 3: DATA PREPARATION PHASE

In Data Preparation Phase, final dataset is prepared from the initial data. Data which is wrong is removed. The appropriate data which is best for analysis is used and other data is dropped. Outliers are treated in this phase. Outlier treatment is done using IQR (Inter Quartile Range) technique.

IQR :

Step 1: Put the numbers in order.

Step 2: Find the median

Step 3: Place parentheses around the numbers above and below the median.

Step 4: Find Q1 and Q3.
Q3 = 75th percentile of data and Q1 = 25th percentile of data.

Step 5: Subtract Q1 from Q3 to find the interquartile range.
IQR = Q3-Q1

Step 6: Data values is defined as an outlier if data is
Lesser than Q1 - 1.5(IQR); or greater than Q3 + 1.5*(IQR)*

STEP 4: MODELLING PHASE

In Modelling Phase, various modelling techniques like CART, C4.5, Random Forest, SVM, Linear Regression are implemented and one or more technique which gives the optimal result is considered. The requirements of a data mining technique used during the modelling phase may cause the process to loop back to the Data Preparation Phase, with the goal of improving data quality.
In this project, C4.5 model will be used to classify applicants in high, medium or low risk.
The C4.5 algorithms recursively visits each decision node, selecting the optimal split, until no further splits are possible.
a. The C4.5 algorithm is not restricted to binary splits.
b. C4.5 produces a tree of more variable shape.
c. For categorical attributes, C4.5 by default produces a separate branch for each value of the categorical attribute.
The C4.5 algorithm uses the concept of information gain or entropy reduction to select the optimal split. Suppose that we have a variable X whose k possible values have probabilities p1, p2 , . . . , pk . The answer is called the entropy of X and is defined as

$$H(X) = -\sum_{j} p_j \log_2(p_j)$$

MIS 637 Midterm
Data Analytics and Machine Learning

October 17, 2019
School of Business
Stevens Institute of Technology

Professor M. Daneshmand

Student Name: SRUJAN JADHAV

C4.5 uses this concept of entropy as follows. Suppose that we have a candidate split *S*, which partitions the training data set *T* into several subsets, T_1, T_2, . . . , T_k . The mean information requirement can then be calculated as the weighted sum of the entropies for the individual subsets, as follows:

$$H_S(T) = \sum_{i=1}^{k} P_i H_S(T_i)$$

where, *Pi* represents the proportion of records in subset *i*. We may then define our *information gain* to be gain(*S*) = *H* (*T*) − *HS* (*T*), that is, the increase in information produced by partitioning the training data *T* according to this candidate split *S*. At each decision node, C4.5 chooses the optimal split to be the split that has the greatest information gain, gain(*S*).

STEP 5: EVALUATION PHASE

In Evaluation Phase, the analyst evaluates the model designed in the modelling phase. They check the quality and effectiveness of the model before deploying it. They also check whether the model meets the requirements which were finalized in the first phase of CRISP-DM process.

In this project, the model should accurately classify the users into ‘High risk applicant’, ‘Medium risk applicant’ and ‘Low risk applicant’. Once the model has great accuracy it is further passed on to the deployment phase.

STEP 6: DEPLOYMENT PHASE

In Deployment Phase, the models created are implemented, reports are generated and parallel mining process can be carried out in different department. Two types of deployments are: Simple and Complex deployment.

In this Project, in Simple deployment, report is generated to check whether the applicant is high risk or average risk or low risk.

In Complex deployment, parallel mining is carried out under real world data. The model should be deployed in such a way that it is easy to upscale or downscale the features in the future. Hence the model deployed will give the classification of the applicant