MIS 637 A Midterm
Data Analytics & Machine Learning

March 22, 2023
12:00 Noon to 2:30 PM

Stevens Institute of Technology

Professor M. Daneshmand

Student Name…Poorvi Raut………………………………

A mortgage company likes to be able to decide on 2 interest rates for new loan applicants as follows: interest rate of 4% for “high risk” applicants, and interest rate of 2.5% for “low risk” applicants. You are being asked to lead 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. 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: In order to deal with such data mining project, we need to follow the Cross Industry Standard Process for Data Mining/ Analytics (CRISP-DM) Model. It is a comprehensive process model to carrying out data mining projects. The process model is independent of both the industry sector and the technology used. The CRISP-DM process model aims to make large data mining projects, less costly, more reliable, more repeatable, more manageable, and faster.

The CRISP-DM reference model for data mining provides an overview of the life cycle of a data mining project. It contains the phases of a project, their respective tasks, and their outputs.

There are 6 phases of CRISP-DM model which are listed below:

1. Business Understanding Phase
2. Data Understanding Phase
3. Data Preparation Phase
4. Modelling Phase
5. Evaluation Phase
6. Deployment Phase

We will provide overview of each phase and also how the above data mining project can be formulated:

1. **Business Understanding Phase**

   Here the discussion is about initial business understanding and selection of type of requirements in terms of business and research. This phase focuses on understanding objectives and requirements from business perspective and then converting this knowledge into data mining problems and basic project plan to achieve these objectives.

   For our data mining project , We need to choose the most appropriate plan for new loan application based on the information provided. There are two interest rates provided for new customers based on category.

   Interest rate 4%: high risk
   Interest rate 2.5%: low risk

As result , We determine the objective of project which should be to predict what type of loan should be offered to the customer.

2. **Data Understanding Phase**

   This phase starts with initial data collection and proceeds with activities in order to get familiar with data, identify data quality problems and discover first insights of data and detecting interesting patterns to provide hypothesis for hidden information.

   Here the steps include:

   - Collect the data.
   - Assess data quality.
   - Perform Exploratory Data Analysis (EDA)

Before classifying the customer as high risk or low risk based on interest rate provided, we need to check their credit history and determine which attributes are of high importance when providing loan.

Some key attributes include:

- Purpose of loan required.
- Customer’s Age
- Employment History
- Source of income
- Savings and assets held.
- Tax statements
- Credit Score
- Any previous loan and repayment history

Data needs to be free of any outliers that will affect the assessment.

Exploratory Data Analysis is required to understand data quality and remove any missing values, outliers in data, summarizing data, replacing the missing values with mean, mode, median , displaying frequency, Plotting the data using scatter plots or histograms to get data insights.

3. Data Preparation Phase

The data preparation phase covers all activities to construct the final dataset (data that will be fed into the modeling techniques) from the initial raw data. Data preparation tasks are likely to be performed multiple times, and not in any prescribed order. Tasks include table, record, and attribute selection, data cleaning, construction of new attributes, and transformation of data for modeling tools.

Cleanse, prepare, and transform dataset.

Our data must also be cleaned up of incorrect information. This involves handling outliers. IQR method (Inter Quartile Range) is used to deal with outliers.
Calculate Lower Quartile (Q1), Median (Q2= 50th percentile), Upper Quartile(Q3) based on where the data is odd or even set of values.
IQR=Q3-Q1 (Q3 is 75th percentile of data and Q1 is 25th percentile)

Data values are defined as outliers if:

Q1-(1.5*IQR) lower than Q1
Q3+(1.5*IQR) upper than Q3

4. Modelling Phase

This is Modelling Phase where analysis of which model to be used and applied to solve the data mining problem. Depends on problem type various techniques are applied to solve the problem like clustering, classification, decision tree algorithms, Artificial Neural Networks.
We calibrate model settings to optimize results.
If required, additional data preparation techniques can be used.

For our data mining problem:

- We select the best modelling technique to classify our customer in high risk/low risk based on interest rate .We therefore use C4.5 Decision tree algorithm .
- C4.5 does not restrict itself to binary splits like CART algorithm. It inherently employs single pass pruning process to mitigate overfitting and works with both discrete and continuous data. Also, it can handle well the issue of incomplete data.
- C4.5 works with concept of information gain and entropy to calculate impurity of cluster. The higher the entropy of cluster the more information is needed to describe the cluster.
- To determine optimal split, say we have variable X has k values with probabilities p1,p2…pk

  For variables with several outcomes use weighted sum of log(p’s)

Entropy(X): =

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

Information Gain:

Suppose we have a candidate split S, which partitions the data into k subsets: T1, T2, …Tk
Weighted sum of the entropies: Sum Pi times
Hs (Ti) = Hs(T), where Pi is the proportion of records in subset i
gain(S) = H(T) – HS(T)
Represents increase in information by partitioning training data *T* according to candidate split *S.*
For each candidate split, C4.5 chooses to split that has maximum information gain, gain(S).

5. Evaluation Phase

Evaluation phase will help the analysts to more thoroughly evaluate the model and review the steps executed to construct the model in order to achieve the business objectives. A key objective is to determine that if any important business issue has been considered adequately or not. At the end of this phase a decision on the use of data mining results should be reached.

For our data mining problem we check the effectiveness of C4.5 Algorithm that we used in modelling phase to classify customer based on high risk/low risk. The more precise and accurate our model is, the better it is.

We check accuracy of out model by calculating metrics like confusion matrix, accuracy , error rate.
Confusion matrix is a table used to define performance of classification algorithm.
The confusion matrix consists of four basic characteristics (numbers) that are used to define the measurement metrics of the classifier. These four numbers are:

TP (True Positive), TN (True Negative), FP(False Positive),FN(False Negative)

Performance metrics of an algorithm are accuracy, precision, recall, and F1 score, which are calculated based on the above-stated TP, TN, FP, and FN.
Accuracy of an algorithm is represented as the ratio of correctly classified customers (TP+TN) to the total number of customers (TP+TN+FP+FN).
Precision of an algorithm is represented as the ratio of correctly classified customers with the loan criteria (*TP*) to the total customers predicted to have the risk (*TP*+*FP*).
Recall metric is defined as the ratio of correctly classified customers (*TP*) divided by total number of customers with risk**.**
Error rate= 1- Accuracy

6. Deployment Phase

Here we make use of the model created in modelling phase and evaluated in evaluation phase . In our data mining problem, we use C4.5
Simple Deployment: Generate Report
Complex Deployment: Repeatable data mining effort.

Below are steps before deploying our model:

- We should monitor our model under human supervision for at least initial training instances on real data before we automate the entire process.
- Then we need to deploy our model in a way that is easier to update changes in future, without having to touch the model.
- Deploying the trained model will also help us in classifying the loan customer as ‘High risk or ‘Low risk’.