Dublin Business School

DATA MINING ASSIGNMENT CA02 (B9DA103)
NAME: KINJAL MARU
STUDENT ID: 10391312
COURSE:MASTERS IN DATA ANALYTICS(BATCH A)

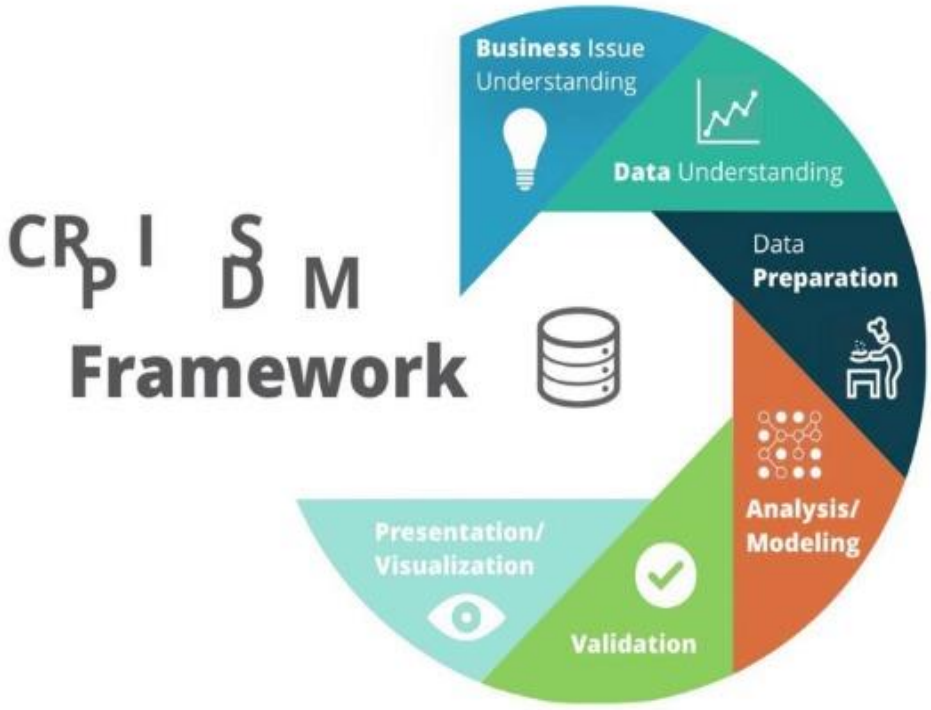
Group Project :-
Bharat Jethwani(10519364)
Kinjal Maru(10391312)
Prasad Tambe(10515513)

INDIVIDUAL CONTRIBUTION/LESSONS LEARNED REPORT :

Comparison of various Car features to predict Car Prices

The car industry has become more and more competitive every year and has grown on a global scale. Therefore, it is necessary to set an accurate price for both customers and manufacturers in this competitive car market. Customers and manufacturers are confused about the price of the car to buy or sell. Consequently, customers and manufacturers try to seek some advice from auto-dealers, car magazines, or the website on the internet. However, this information takes a long time and might confuse the customers in the market.

So in order to predict the price of a car we have implemented CRISP-DM methodology, in our automobile price dataset. CRISP-DM is a data mining application or technique that allows you to perform a data mining task, or helps you with a blueprint.



The method comprises of six phases namely: business understanding, data understanding, data preparation, modeling, evaluation, and deployment that enables the association to offer a good roadmap for the data mining process to follow.

Business understanding:

This contains the business development goal, the issue description, and the main steps toward achieving a project plan. Our data contains unique features of the car which can help to predict the price of the car.

In this section after brainstorming, with the given data we eventually came up above objectives to achieve. As it is important to check what actually affects the price of the car

To check if manufacture affects the price of the car. Also, can horsepower alone make an impact on the price of a car? Whether other features like highway-mpg, stroke, peak-rpm, wheel-base, curd-weight affects the car price prediction? Also if width or height any one of them can make a difference in price prediction? These are the basic objective of our project.

Data Understanding:

It deals with project initialization, where data is first obtained and various activities are carried out to explain the data.

Our data consists of 26 features that describe different parts of the car with 205 unique values and 59 missing values in it. This data collection consists of three categories of entities:

1) With various features classification of car prediction

2) The insurance risk rating which is given to the car model

3) The consistent accidents in the operation with respect to other cars.

The second-ranking represents the level to which the car is riskier than its price suggests. Initially, cars are given a symbol of the risk factor associated with their size. The third element is the overall average damage premium per every vehicle year covered. This amount is common for all cars under a similar size category (small two-door vehicles, station wagons, sports/specialty, etc.) and reflects the average annual loss per vehicle.

We have learned to make even 3D visualizations from RapidMiner and made different types of visualization to understand the data well as which feature is affecting the label variable in Tableau.

Data Preparation:

Multiple steps are involved in transforming data to feed into modeling tools.

In this section, I have learned about how can we process the data in RapidMiner with the inbuilt model. We have inserted data in the RapidMiner and tried to process the same. We have used impute missing values function to replace missing values using the KNN technique. We have normalized our data so that we do not get bias and variance and model get better results. The feature selection technique enables the algorithm to train faster as well as reduces complexity. After trying both forward and backward selection techniques,we found that optimization selection was best amongst them.

Modeling:

From given data model runs on the training dataset and it is checked on the test dataset to interpret the accuracy of data mining model prediction. I have learned from this section that we may get better results even if we don't rely on the auto model in RapidMiner. As auto model suggested the Random forest as the best model for our dataset. However, after investigation we came to know that we are getting much better results using the KNN method. To execute the model, we have split the data into 70:30 ratio and used the KNN model.

Evaluation:

Data analyst verifies whether the model meets the business goals and tests the model on real-world applications. Here I have checked both the Model's performance, i.e. one which auto model suggested and the one which we have found out using trial and error of different method and feature selection methods. After Evaluation of the results of different models we decided to go with KNN model as we got better results with this model. We have found seven features which are important for predicting the price of the car namely width, curb-weight, engine-size, horsepower, peak-rpm, make.

Deployment:

Project implementation can be accomplished by producing reports based on the findings and discussing the strategies used in the data mining process with the client. I have learned about how the deployment is done in RapidMiner.There are below parts which needed to deploy a model.

1) You need to prepare the data with TurboPrep
2) You need to build the model with Auto Model
3) In the Deployments View ,Models can be deployed

We wanted to explore the deployment of a rapid miner so we did the part of deployment by deploying a random forest model to our local repository with the performance vector.

Conclusion:

From this project, I have learned that there are few features that affected the price of the car like horsepower, width, and maker's name. However, we can't predict based on just this much information in the real-life car prediction goal. So we can deduce that we can incorporate Prediction Models using similar approaches, technology and tools like Rapid Miner and Tableau.