Machine Learning Engineer Nano degree

Capstone Project Proposal

Prediction of Bike Buyers

Sanke Sneha

February 21st

Proposal:

Prediction of Bike Buyers

Domain Background:

In this project, I am working on prediction of bike buyers. Transportation has been part of our life as long as humanity exists. The most common road vehicle for transportation is automobile. Automobile includes cars, motorbikes etc.

Bikes are the affordable vehicles for all kinds of people. As the need of transportation increases the usage of bikes also increased which results in the sudden rise of automobile industries. The automotive industry continues to face set of challenges. Most manufacturing operations in automotive industries are still largely dependent on experiences-based human decisions. I am highly interested to apply Machine Learning in the automotive industry to make a remarkable ability to bring out hidden relationships among datasets and make predictions.

The related academic work is found at https://towardsdatascience.com/predicting-no-of-bike-share-users-machine-learning-data-visualization-project-using-r-71bc1b9a7495 . It is somewhat similar to my project and I got inspired by this to apply machine learning in automotive industry as well. In this article it is taken as regression task and my task follows classification.

Problem Statement:

My main aim is to predict the bike buyers. For doing this I selected my private bike buyer dataset

Bike Buyer Dataset

So the goal is to predict the bike buyers. Here I am using classification models to find the f-score of each model and select the best model with high f-score to predict the bike buyer or not. Here the input is training data that we took and output will be whether bike buyer or not.

Datasets and inputs:

Number of attributes: 13(11-predictive, 1-non predictive, 1-goal field)
Number of instances: 6998

The dataset that I am working is my private data. The dataset is quite interesting because it is a good mixture of categorical and numerical attributes. Each feature has certain effect on the solution we are going to predict. Some customers who are married and having children are more likely to buy cars rather than bikes and if they already having cars are also not that much willing to buy. If the commute distance is long they are certainly willing to buy. Sometimes female genders are not willing to buy or vice versa. According to my intuition yearly income, gender, cars, children, commute distance are more important features. The dataset I am using is unbalanced because among 7000 customers [5997] are 'non-buyers' and [998] are 'buyers'.

An example from the dataset is the person whose gender is male, married having 5 children and 1 car with yearly income 160000 is not willing to buy the bike. The person who is male, single with 1 car and yearly income 30000 is willing to buy the bike.

Attributes:

ID

Marital Status: Married or Single

Gender: Female or male

Yearly Income: range of income that a customer is having

Children: number of children of the customer

Education: Qualification of the customer

Occupation: occupation of the customer

Home Owner: having a own house or not

Cars: Number of cars

Commute Distance: distance from home to work place

Region: Place of the customer

Age: age of the customer

Bike Buyer: yes or no

Solution Statement:

Here I am trying to predict the bike buyers from the selected dataset. For prediction we will use the different classification models svm, RandomForest and decision tree. Then we will find the f-score, accuracy, precision of each model. I will use the model with high f-score to make best predictions. I explore the dataset using read_csv and for visualization which helps me to better understand the data I use matplotlib.pyplot.

Benchmark Model:

Benchmark model is a model which we will take as reference and achieve the best than the benchmark model. In our project, we will take knn classifier as benchmark model. Using knn classifier we will achieve fscore 0.2. Now we will try and achieve the better fscore than the benchmark model.

Evaluation Metrics:

I want to use accuracy, precision and f-score as evaluation metric for prediction of bike buyer. Here I am predicting f-score for selected models. Here we will select a model whose f-score is greater than all the other models and we treat it as the best.

Project Design:

Project is composed of different steps as follows.

Pre-Processing:

The first task is to read the data after that clean the data that is removing the data or treatment for missing values or removing duplicates etc.

After the data exploration, we split the data into training and testing sets. After splitting of data to change categorical data into numerical data we will use LableEncoding and OneHotEncoding.

After that we will visualize the data to get some deep insights for interpretation. Later we will apply different Classification models and then predicts the accuracy, precision and f-score of each model.

Training the data:

Here I will use the classification models Decision tree, svm and RandomForest. After training the data we will test each model with testing data. After that we will find out the f-score for each model.

Finally I will declare the model with high f-score as best model for the prediction of bike buyer.