Machine Learning Engineer Nanodegree

Capstone project proposal

Pasupula Divya Sai

February 3rd,2019

Proposal:

Car Evaluation dataset

Domain background:

History:

The automotive industry is the pillar industry of the national economy, and millions of people are closely related to it. For most people, buying a car is to buying a house outside in addition to a maximum consumption, can better reflect consumer demand and the real market behaviour.

Cars are essentially part of our everyday lives. There are different types of cars as produced by different manufacturers; therefore the buyer has a choice to make. The choice buyers or drivers have mostly depends on the price, safety, and how luxurious or spacious the car is.

when an individual consider of buying a car, there are many aspects that could impact his/her choice on which kind of car he/she is interested in. There are different selection criteria for buying a car such as prize, maintenance, comfort, and safety precautions, etc. Here I will be applying various machine learning classification algorithms to the car evaluation dataset.

Therefore, a reasonable evaluation method is equally important for car consumers(buyers) and producers. It can not only reduce the burden on dealers, but also increase sales. In addition, it plays a strategic role, can improve customer service levels in a highly competitive market environment.

So,I am going to evaluation this project

One recent research paper based on the car evaluation reference link:

https://www.ijcaonline.org/archives/volume172/number9/28279-2017915205

Problem Statement:

The main aim of my project is to find out which of the classifier best suited for the dataset.For this I selected the data set compiled from a wide range of sources So,My goal is to evaluate the car dataset by below mentioned classifiers.Here I am using the classification models to find the accuracy

of each model and select the best model which will have high accuracy.The model created with the training dataset has been evaluated with the standard metrics such as accuracy, f score.

The experiment will be carried out using some classifier models, namely;k nearest neighbour,logistic regression, decision tree classifiers. This is in view to finding out which of the classifier best suits the dataset in terms of classifying the pre-processed data, trained data, testing, and making prediction using the model obtained from the training process.

Datasets and Inputs:

The dataset used in this study which is a collection of the records on specific attributes on cars donated by Marco Bohanec in 1997 was obtained from the UCI dataset repository. The car evaluation dataset as described in the UCI dataset repository was derived from simple hierarchical decision, and is categorized descriptively in table 1

Dataset characteristics	Multivariate	Number of Instances present	**1728**
Attribute characteristics	Categorical	Number of Attributes	**6**
Task to be done	Classification	Missing Attribute values	**None**

The Car Evaluation Database contains examples with the structural information removed, i.e., directly relates CAR to the six input attributes namely buying, maint, doors, persons, lug_boot, safety.

This data is downloaded from https://archive.ics.uci.edu/ml/datasets/Car+Evaluation

Number of Instances: 1728

Number of Attributes: 6

Attribute Names:

Buying,maint,doors,persons,lug-boot,safety.

Attribute Values:

buying v-high, high, med, low

maint v-high, high, med, low

doors 2, 3, 4, 5-more

persons 2, 4, more

lug_boot small, med, big

safety low, med, high

Solution Statement:

Here, I trying to predict the best classifier suited for the selected data set.For predicting this I am going to use the different classification models like logistic regression,k nearest neighbour,decision tree.Then,I will find the accuracy score for each classification model. I will explore the data set with matplotlib.pyplot libraries, Practicing with scikit learn, pandas, numpy in this project.By using visualization I can better the understand the solutions.

Benchmark model:

This step will be important because compare your final model with some of them and see if it got better, same or worse. Here accuracy score will be compared between the models and select the best one.However I target to reach greater than 85% in my attempt to solving this project.

Evaluation metrics:

I want to use accuracy score as my evaluation metric for predicting the best classifier for my dataset . Here I am predicting the accuracy score for the selected models. In multilabel classification, this function computes subset accuracy: the set of labels predicted for a sample must exactly match the corresponding set of labels in y_true. Here the model with the highesht value it is selected as the best model.

Project Design:

My experiment will be carried out using three classifier models, namely:logistic regression,k nearest neighbour,decision tree classifiers. This is in view to finding out which of the classifier best suits the dataset in terms of classifying the pre-processed data, trained data, testing, and making prediction using the model obtained from the training process.

The project is composed of different steps as follows:

Pre-processing: First task is to read the dataset and perform visualizations on it to get some insights about the data and attributes. After reading the data,

The data as obtained from the UCI dataset repository have to be cleaned and to ensure that it is in the standard quality before the model creation is initiated.So clean the data i.e. removing unwanted data or replacing null values with some constant values or removing duplicates if any. Then finding the correlation for each features with the target variable

After Data Exploration, I want to split the total data into training, validation and testing data and normalize the data to make it suitable for free from .Then applying the different classifying models and then predicting the accuracy score to the selected models

First step in training: First, I want to choose a Benchmark model which will at least gives testing accuracy score around 60% accuracy score. Second step in training: I want to apply classification models of my own and use on the data. I want to apply K nearest neighbour and Logistic Regression model then find the Accuracy score for both the models.Later with some metrics like gridsearchCV,cross validation method.Then come to a conclusion to some accuracy better than benchmark.

Finally, I will declare the model which highest accuracy score on both training and testing data sets concluded as the best model for car evaluation dataset and perform well in my attempt.