Machine Learning Engineer Nanodegree

Capstone Proposal

Shaik Arshiya Yasmine

February 22nd, 2019

Proposal

Car Evaluation

Domain Background:

Essentially these are the means for assessing a car. By utilizing this we can pass judgment on whether the vehicle is commendable or not.

Stage I: Research the Car's VIN. Each vehicle has a novel vehicle distinguishing proof number (VIN).

Stage II: Informal Inspection.

Stage III: Test Drive.

Stage IV: Professional Inspection

While coming to specialized viewpoints these are the past use Vehicle Evaluation Database was gotten from a basic various levelled choice model initially produced for the exhibition of DEX, M. Bohanec,V. Rajkovic: Expert framework for basic leadership. Sistemica 1(1), pp. 145-157, 1990.).

M. Bohanec and V. Rajkovic: Knowledge obtaining and clarification for multi-trait basic leadership. In eighth Intl Workshop on Expert Frameworks and their Applications, Avignon, France. pages 59-78, 1988.

Problem Statement:

The issue explanation is to assess a vehicle dependent on an objective idea (CAR), the model incorporates three halfway ideas: PRICE, TECH, COMFORT

The objective variable is CAR (car agreeableness) which implies whether the vehicle is commendable or not. This should be possible by halfway ideas which I described below.

The model assesses cars as indicated by the accompanying idea structure:

CAR -car acceptability

- PRICE- overall price
  - buying- buying price
  - maint -price of the maintenance
- TECH -technical characteristics
- COMFORT-comfort
  - doors-number of doors
  - persons-capacity in terms of persons to carry
  - lug_boot- the size of luggage boot
  - safety stimated safety of the car

By utilizing these intermediate ideas we will make judgment about that car. Based on these ideas we will get an end whether the vehicle is great or awful.

Datasets and Inputs:

Here the dataset will have categorical values. There will be aggregate of 1728 instances, and the all out number of attributes is 6. The dataset here is a multivariate.

Dataset reference: https://www.kaggle.com/elikplim/car-evaluation-data-set#car_evaluation.csv

Attribute information:

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

Class Distribution (number of instances per class):

class	N	N[%]
unacc	1210	(70.023 %)
acc	384	(22.222 %)
good	69	(3.993 %)
v-good	65	(3.762 %)

This depicts the distribution of target class(CAR acceptability).The absolute 1728 instances are isolated into 4 classes as referenced previously.

Here I am utilizing a Multivariate dataset.

Solution Statement:

Here I am trying to foresee the result from the chose information (dataset). For doing as such we need to utilize distinctive classification models. At that point we will discover the accuracy score for every classification model. I inspect the dataset with read_csv, and matplotlib.pyplot libraries in this undertaking. By utilizing visualization helps me to more readily comprehend the solution.

Benchmark model:

This step will be essential because we compare the last model and some of the other models and check whether it showed signs of improvement, same or more terrible. Here Accuracy score will be compared between the models and the best one will be selected.

Here I might want to utilize straightforward model, for example, basic logistic regression to get a gauge score for my dataset.

Evaluation Metrics:

Here I will utilize accuracy score as evaluation metric. I will anticipate the accuracy score for the chosen model. The model with high accuracy will be the best model out of the picked models.

Accuracy can't be a reasonable foundation to assess imbalanced classification. So I checked for f1-score also. The model with better f1 score will be the best model to assess the vehicle.

Project Design:

The project consists of the below steps:

Pre-processing:

First thing is to read the dataset and perform some visualization on it to get a few insights about the data. After that, clean the data i.e., removing undesirable information or substituting null values with some constant values or removing duplicates.

After the exploration of dataset, I will split the dataset into training set and testing set. And then apply the classification models and predict the accuracy score to the chosen models.

Training and Testing:

I need to apply classification models of my choice and train them on the data. I need to apply logistic regression, KNN classifier and random forest.

At that point, I will find the accuracy score for the previously mentioned models. For this I will initially train the models with the training dataset, and afterward continue testing with the testing dataset that I split previously.

At long last, I will proclaim the model which has the most noteworthy precision score out of all the picked calculations and announce it as the best one.