Machine Learning Engineer Nanodegree

Capstone Project Proposal

Suthraye Swamynath Rao

February 8th 2019

Proposal:

Car Evaluation

Domain Background:

History:

Basically these are the steps for evaluating a car.by using this we can judge whether the car is worthy or not.

Step 1: Research the Car's VIN. Every car has a unique vehicle identification number (VIN). ...

Step 2: Informal Inspection. ...

Step 3: Test Drive. ...

Step 4: Professional Inspection

While coming to technical aspects these are the past usage

Car Evaluation Database was derived from a simple hierarchical decision model originally developed for the demonstration of DEX, M. Bohanec, V. Rajkovic: Expert system for decision making. Sistemica 1(1), pp. 145-157, 1990.).

M. Bohanec and V. Rajkovic: Knowledge acquisition and explanation for multi-attribute decision making. In 8th Intl Workshop on Expert Systems and their Applications, Avignon, France. pages 59-78, 1988.

Database reference:https://archive.ics.uci.edu/ml/datasets/Car+Evaluation

Problem Statement:

The problem statement is To evaluate a car based on a target concept (CAR), the model includes three intermediate concepts: PRICE, TECH, COMFORT

The target variable is CAR(car acceptability) which means whether the car is worthy or not.This can be done by intermediate concepts which I discribed below.By

The model evaluates

cars according to the following concept 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	estimated safety of the car

.By using this intermediate concepts we will make judgement about that car.Based on these concepts we will get a conclusion whether the car is good or bad.

Datasets and input parameters:

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

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 describes the distribution of target class(CAR acceptabilty)..The total 1728instances are divided into 4 classes as mentioned above.
Here I am using a Multivariate dataset.

Solution Statement:

Here I am trying to predict the outcome from the selected data. For doing so we want to use different classification models. Then we will find the accuracy score for each classification model. I explore the dataset with read_cv, and matplotlib.pyplot libraries in this project. By using visualization helps me to better understand the solution.

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 compare between the models and select the best one.
Here I would like to use simple model such as simple logistic regression to get a baseline score for my dataset

Evaluation Metrics:

Here I will use accuracy score as evaluation metric. I will be predicting the accuracy score for the selected model. The model with high accuracy score will be the best model out of the chosen models.
accuracy can't be a fair criterion to evaluate unbalanced classification, so I checked for 'f1-score also.The model with better f1 score will be the best model to evaluate the car.

Project Design:

The project is composed of the following steps:

Pre-processing:

First task is to read the dataset and perform visualization on it to get some insights about the data. After reading the data clean the data that is removing unwanted data or replacing null values with some constant values or removing duplicates

After data exploration I will split the data into training set and testing set. Then applying the classification models and predicting the accuracy score to the selecting models.

Training and Testing the data:

I want to apply classification models of my own and use them on the data. I want to apply logistic regression, KNN classifier and random forest.

Then I will find the accuracy score for the above mentioned models. For this I will first train the algorithms with the training data, and then carry on testing with the testing data that I split before

Finally, I will declare the model which has the highest accuracy score out of all the chosen algorithms and declare it as the best one.