Machine Learning Engineer Nanodegree

Capstone Proposal

Ramakrishna

June 23rd,2018

USING CLASSIFICATION TO EVALUATE CARS

Domain Background

In this project I am going to take a car evaluation dataset which is from Automobile Marketing Domain. In today's Automobile marketing domain every company or seller wants to sell more and more products to customers. In order to make that happen the sellers should know what type of automobiles the customers prefer based on some features like safety, maintainence etc. The Automobile marketing domain has drastically changed from the past few decades with the invention of more and more new models. In order to survive in this domain the company should know the customer requirements and design their models according to them.

Problem Statement

Here the problem is Car Acceptability by the customers i.e., based on the some features of a car we have to tell that a new car with the same features is acceptable by the customers or not. This can be handled by using classification algorithms like logistic regression, KNN, Random Forests etc. We have to calculate accuracy and F-score for our training set and check it works fine for our testing data among different models.

Datasets and Inputs

I have taken Car Evaluation Dataset from UCI machine learning repository (https://archive.ics.uci.edu/ml/datasets/Car+Evaluation) which contains 1728 records and 6 attributes. The attributes with different values are as follows:

Buying(buying price): vhigh, high, med, low.
maint(maintainence price): vhigh, high, med, low.
doors(No of Doors the car have): 2, 3, 4, 5more.
persons(No of persons that car can fit): 2, 4, more.
lug_boot(luggage boot size): small, med, big.
safety(safety level): low, med, high.

Based on the above attributes we have to give car acceptability labels as unacc, acc, good, vgood. I considered these attributes because the customers first see the buying price of the car and how many members can fit in the car and the maintainence cost it will take and mainly safety levels of the car which are very important compared to other features.

Customers also see the company of the car, color of the car etc but the above attributes plays an important role while comparing with other attributes and they itself will be sufficient to tell the car acceptability values.

Solution Statement

The above problem can be solved using multivariate classification techniques like Logistic Regression, KNN, Random Forests etc.. because the dataset is small and we have only six attributes in our dataset. We have to calculate accuracy for each of the selected algorithms and we can go with F-score in order to tell which algorithm performs well with our dataset. We have to make sure that our model does not overfit our testing set and make it more and more accurate for our testing set.

Benchmark Model

First I will use logistic regression as my benchmark model and calculate it's accuracy and F-score later I will use another model and compare it's accuracy and F-score with benchmark model and I finally decide which model to go with and I will make that model as my solution model.

Evaluation Metrics

I will use Accuracy as my evaluation metric for both benchmark and solution models. In order to calculate accuracy first I divide my data into training and testing sets and then by using training set I will fit the training data into my model and I will predict the labels using X-test and then calculate accuracy score using both test sets.

Project Design

Firstly I will change my categorical data with integer data using labelEncoder and transform my categorical data into integers. Later I will select three models like logistic Regression, KNN and Random Forests. First I go with logistic regression and calculate accuracy of my model. With remaining two models also I will fit my data and calculate accuracy score and check whether the model is overfitting with my data or not by plotting the learning curves for my models and if it is overfitting means I calculate F-score and tell whether my model is good or not for my data. I also work with different parameters of that model to get better results. Finally based on the accuracy and F-score and with parameter tuning I will pick the model with high accuracy and F-score as my solution model.