MACHINE LEARNING ENGINEER NANODEGREE

CAPSTONE PROJECT

CAR EVALUATION

1.Definition

Project Overview:

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

The main aim of this project is to predict the best suited classifier for car evaluation dataset.

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.

Metrics:

I want to use accuracy score as my evaluation metric for predicting the best classifier for my dataset.

Accuracy is a common metric for binary classifiers,it takes into account both true positives and true negatives with equal weight.

Accuracy measures how often the classifier makes the correct prediction. It's the ratio of the number of correct predictions to the total number of predictions (the number of test data points).

$$\text{Accuracy} = \frac{\text{true positives} + \text{true negatives}}{\text{dataset size}}$$

It would seem that using **accuracy** as a metric for evaluating a particular model's performace would be appropriate.

Further i would like to use f-score if necessary.

F score is a measure of test's accuracy.It consider both precision and recall of the test to compare the score. ** Precision ** tells us what proportion of messages we classified as spam, actually were spam.

It is a ratio of true positives(words classified as spam, and which are actually spam) to all positives(all words classified as spam, irrespective of whether that was the correct classificatio), in other words it is the ratio of

`[True Positives/(True Positives + False Positives)]`

** Recall(sensitivity)** tells us what proportion of messages that actually were spam were classified by us as spam.

It is a ratio of true positives(words classified as spam, and which are actually spam) to all the words that were actually spam, in other words it is the ratio of

`[True Positives/(True Positives + False Negatives)]`

Confusion Matrix : A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known. The confusion matrix itself is relatively simple to understand, but the related terminology can be confusing.

True positives (TP): These are cases in which we predicted yes (they have the disease), and they do have the disease.

True negatives (TN): We predicted no, and they don't have the disease.

False positives (FP): We predicted yes, but they don't actually have the disease. (Also known as a "Type I error.")

False negatives (FN): We predicted no, but they actually do have the disease. (Also known as a "Type II error.")

Reference link: https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/

https://en.wikipedia.org/wiki/F1_score

2.Analysis

The objective of data analysis step is to increase the understandinng of the problem by better understanding the problems of the data.

There are two approaches to describe a given dataset.

Summarizing and Visualizing data.

Data Exploration:

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.

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

There are 1728 instances and 6 attributes in my dataset.

The information about each attribute is explained below:

1.Buying(buying price)

2.Maint(price of maintenance)

3.Doors(number of doors)

4.Persons(capacity in terms of persons to carry)

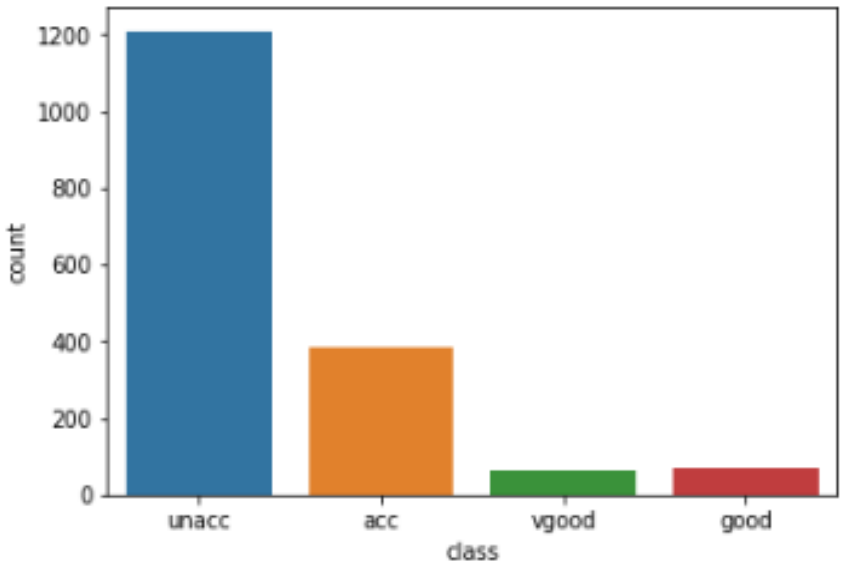
5.Lug_Boot(the size of luggage boot)

6.Safety(estimated safety of the car)

Exploratory Visualization:

For data visualization I import a library called seaborn which is a library for making statistical graphics in Python and matplotlib which is a Python 2D plotting library which produces publication quality figures in a variety of formats across the platforms.

Reference link: https://matplotlib.org



It can be seen from the graph that the result 'class' is unbalanced with larger values of 'unacc'.

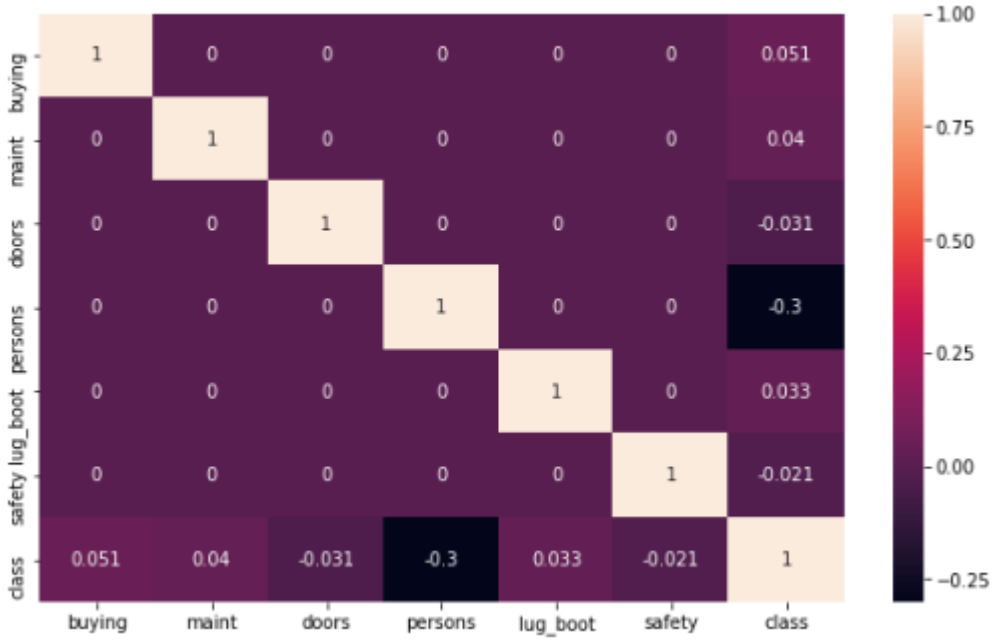
Now this is an unbalanced multiclass classification problem.

Here is a heatmap,the correlation of features of data.A heat map is a two-dimensional representation of data in which

values are represented by colors. A simple heat map provides an immediate visual summary of information. More elaborate heat maps allow the viewer to understand complex data sets.¶

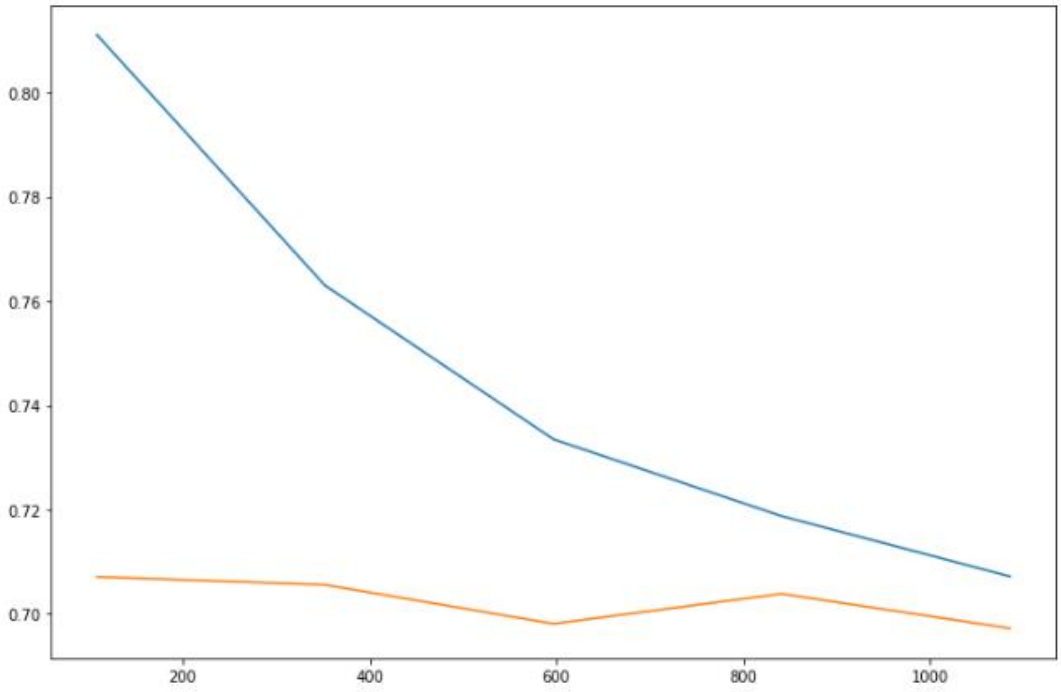
Heatmap of the columns on dataset with each other. It shows Pearson's correlation coefficient of column w.r.t other columns.

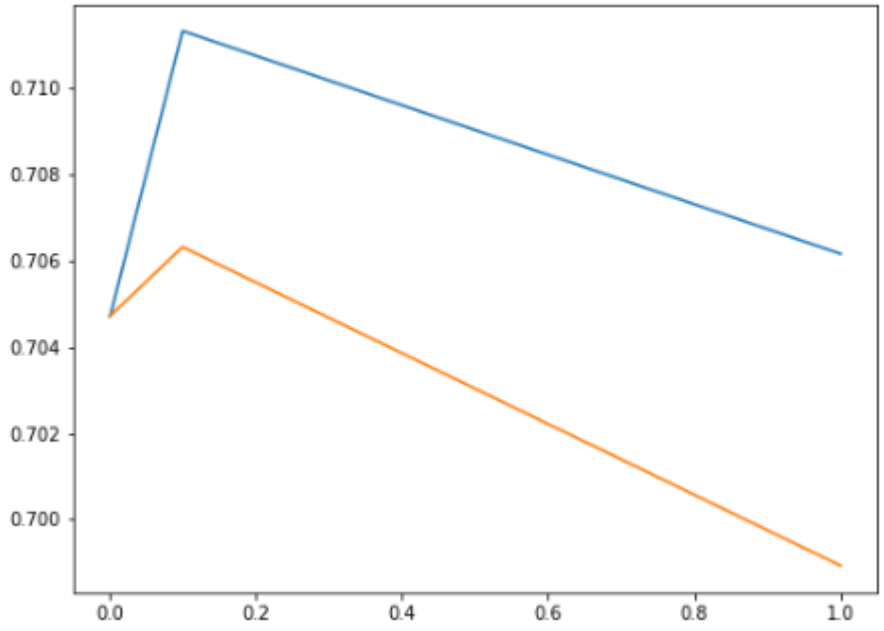
The map is shown below:



Next comes some learning curves,conventionally depicts improvement in perfomance on the vertical axis when there are changes in another parameter.

If we plot the evolution of two errors scores as training sets change we end up with two curves.These are called learning curves.


0.80
0.78
0.76
0.74
0.72
0.70
200
400
600
800
1000


0.710
0.708
0.706
0.704
0.702
0.700
0.0
0.2
0.4
0.6
0.8
1.0

Algorithms and Techniques:

The algorithms which I am going to use are mentioned in my proposal namely Logistic Regression,K nearest neighbour,Decision tree.

Logistic Regression:

Logistic Regression is a technique borrowed by machine learning from the field of statistics.Logistic regression predicts the probability of an outcome that can only have two values .The prediction is based on the use of one or several predictors (numerical and categorical).Moreover logistic regression is a predictive analysis.When selecting the model for the logistic regression analysis, another important consideration is the model fit. However, adding more and more variables to the model can result in overfitting, which reduces the generalizability of the model beyond the data on which the model is fit.I can see two main advantages of logistic regression.The first is you can include more than one explanatory variable (dependent variable) and those can either be dichotomous, ordinal, or continuous. The second is that logistic regression provides a quantified value for the strength of the association adjusting for other variables (removes confounding effects).

Strengths: Logistic regression also performs better (than Naive Bayes) if your features are not conditionally independent. But logistic regression has the advantage over decision trees and SVM of allowing you to update your model as you receive new data, and producing probababities so that

you can measure the confidence level of the model's predictions.

Weaknesses: Logistic regression doesn’t perform well when the feature space is too large and/or there is a large number of categorical features. It also requires you to perform transformations for non-linear features and may be influenced by outliers since it relies on the entire data set.

Parameters: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

Reference: https://www.statisticssolutions.com/what-is-logistic-regression/

https://machinelearningmastery.com/logistic-regression-for-machine-learning/

K nearset neighbour:

The k nearest neighbour algorithm is a simple,easy to implement supervised machine learning algorithm that can be used to solve both classification and regression tasks.

The KNN algorithn assumes that similar things exist in close proximity.KNN has no model other than storing the entire dataset,so there is no learning required.

When KNN is used for classification, the output can be calculated as the class with the highest frequency from the K-most similar instances. Each instance in essence votes for their class and the class with the most votes is taken as the prediction.

Advantages:
The K-Nearest Neighbor (KNN) Classifier is a very simple classifier that works well on basic recognition problems.
Disadvantages:
The main disadvantage of the KNN algorithm is that it is a *lazy learner*, i.e. it does not learn anything from the training data and simply uses the training data itself for classification. To predict the label of a new instance the KNN algorithm will find the *K* closest neighbors to the new instance from the training data, the predicted class label will then be set as the most common label among the *K* closest neighboring points. The main disadvantage of this approach is that the algorithm must compute the distance and sort all the training data at each prediction, which can be slow if there are a large number of training examples. Another disadvantage of this approach is that the algorithm does not learn anything from the training data, which can result in the algorithm not generalizing well and also not being robust to noisy data. Further, changing *K* can change the resulting predicted class label.
Reference link: https://scikit-learn.org/stable/modules/neighbors.html#nearest-neighbors-classification

Decision tree:

Decision tree is a general predictive modelling tool that has application spanning a number of different areas.They are constructed via an algorithmic approach that identifies ways to split a dataset based on different conditions.

Decision trees are a non-parametric supervised learning method used for both classification and regression tasks.

Strengths:It is very easy to understand and interpret.The data for decision trees require minimal preparation.

Weaknesses:Sometimes decision tree may become complex.The outcomes of decisions can be based mainly on your expectations.So this can lead to unrealistic decision trees.

Since a decision tree can handle both numerical and categorical data, it's a good choice of algorithm.

The goal is to create a model that predicts the value of target variable by learning simple decision rules.

https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/ml-decision-tree/tutorial

---

Parameters:

https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

As my application is classification oriented .So, techniques that are used are taken from Classification models.

Benchmark:

Benchmarking requires that experiments be comparable,measurable and reproducable.

This step will be important because compare your final model with some of them and see if it got better, same or worse.

My blind benchmark accuracy will be around 60%.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.

3.Methodology

Data preprocessing:

Before data can be used as input for machine learning algorithms, it should often must be cleaned, formatted, and restructured — this is typically known as preprocessing. Fortunately, for my dataset, there are no invalid or missing entries we must deal with.

It refers to transformations applied to our data before feeding it to the algorithm.

The technique that is used to convert raw data into a clean data set.

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.

Data transformation is a very crucial process in data preprocessing.It involves normalization and aggregation.

Normalization: Normalization makes training less sensitive to the scale of features, so we can better solve for coefficients.

After preprocessed dataset is split into two halves of varying sizes at different times for use as training and testing datasets

for model creation and selection of which of the models performs best.

The data set used for training is mainly a portion from the dataset from which the classifying algorithm used learns the class/result of the model created from each model.

The whole data is divided into training and testing data using train_test_split from sklearn.model_selection.

Implementation:

The implementation process can be split into two main stages.
1.The classifier training stage
2. Tuning the parameters of best defined classifier
Classifier traning stage involves selecting the classifier for model creation with above mentioned algorithms namely
Logistic Regression
K nearest neighbour algorithm
Decision tree
Firstly I tried logistic regression the accuracy was 66.4%,later after regularizing the parameters using cross validation scores,learning curves the accuracy increased to approximately 71%.
Secondly,I tried KNN algorithm with a rapid increase in accuracy i.e., 90%.
Finally i used DecisionTreeClassifier to get better accuracy of 96.5% which crossed my benchmark accuracy.

Tuning the parameters stage involves the parameter of the best defined classifier are tuned using appropriate methods to get best accuracy score.¶

1.After getting the necessary parameters of the best defined classifier the training data is fitted to the Classifier.

2.Now the scores like accuracy score, f1 score etc are obtained from the data using the Classifier.

Refinement:

Since I started my project I have started using basic classification algorithm implemented by knowing about it from various sources.Then I explored for different algorithms for more better results and reached KNN algorithm which was easy to implement.Later I reached tuning the parameters to finally reach a best suited algorithm i.e.,DecisionTree algorithm where I achieved good accuracy as I target to get.I used cross validation score,confusion matrix to make it better for me .I think I made my work more refined going forward than previous.

4.Result

Model Evaluation and Validation:

Reference link: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.make_scorer.html
http://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html

The classification report:

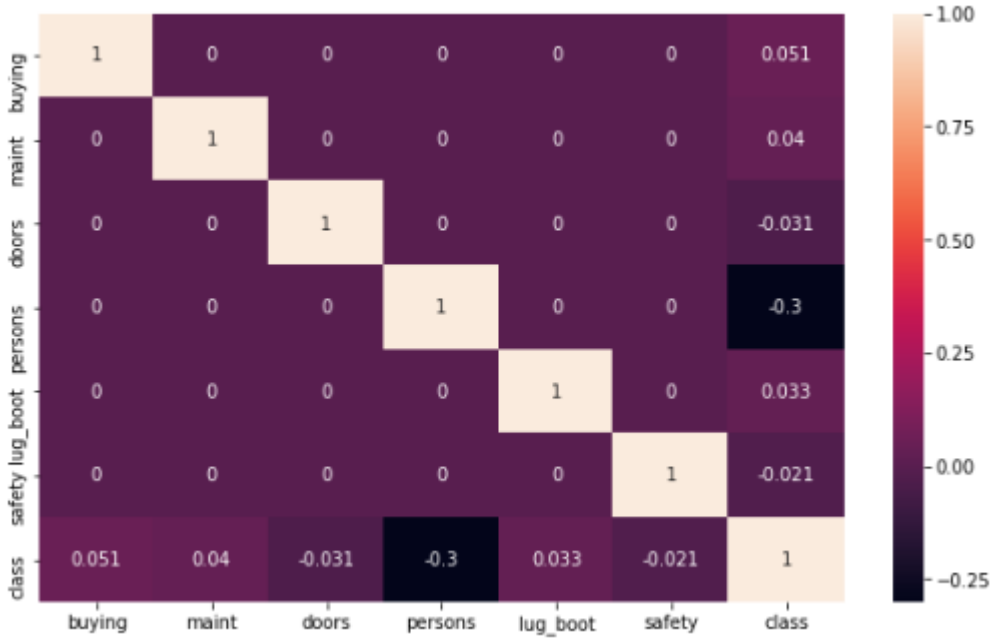
	precision	recall	f1-score	support
0	0.82	0.79	0.80	118
1	0.77	0.53	0.62	19
2	0.93	0.99	0.96	358
3	1.00	0.50	0.67	24

Justification:

Three models used in the empirical car acceptability evaluation were compared and researched, the results showed that all three models have good predictive ability, and decision tree algorithm model showed the best accuracy, sensitivity and specificity, with the best predictive ability, can be very good for car acceptability evaluation. In the six attributes of the car, the safety has the largest influence on car acceptability followed by occupancy, however, consumers is less sensitive to these two factors of the size of the trunk and the number of door. This can help companies make better policy, targeted, improved methods to improve the car's acceptability and consumer satisfaction. The benchmark accuracy is around 60%.The optimised model has obtained accuracy(96.5%) better than my benchmark.So,it is performing well better.It is tell us that Classification Report shows good results of recall, f1-score, and support.

5.Conclusion

Free form Visualisation:



This helped me a lot in visualizing data better.

Heatmap is a two-dimensional representation of infomation with the help of colors.Heatmap of the columns on dataset with each other.It shows Pearson's correlation coefficient of column w.r.t other columns.
Learning curves also helped in better visualization of my data.

Reflection:

1.During my process of doing this project, I learnt how to visualize, and understand the data.
2.I have learnt that Data Cleaning plays crucial part in Exploratory Data Analysis(EDA).
3. Removing the data features that are not necessary in evaluating an model is most important.
4.I have come to know how to use best algorithm in different conditions for the data using appropriate techniques.”sklearn” helped me a lot in knowing a lot about the respective algorithms and their parameters.
5.I am aware of how to tune the parameters to get best score.
6 At last , I have learnt how to grab a data set from machine learning repository and applying techniques on it and to stick to best techniques to get good results.Finally I am glad that I can solve a problem and acquire a solution using machine learning concepts.

Improvement:

In this project, I have evaluated the different classifiers for car evaluation dataset. Based on the customer feedback about the cars used, the model is very appropriate to judge the best car segment as per the requirement of the customer. In future, research can be use more refine technique to give more accuracy and deal with the some other issue like choose the nature of feeling, also assemble the traverse of the testing dataset and can take a gander at the more auto evolution as enormous number of flexible car are available in market. Not simply with compact brand however for other thing we can

perform same investigation.We can further use ensemble methods to get better results I hope to expect.