Machine Learning Engineer Nanodegree

Using Supervised Learning to identify whether a car is worth or not on its with respect to its features.

Geetha Harika Darimireddy

June 26th,2018

Proposal

Domain Background

A car with specific features and specifications is worth or not is very important to a customer. It is not a surprise that the car with particular features is very profitable for a customer. The car is not worth means that may not be safe or may be the maintenance is high and comfort is low. However, this issue become ultra-costly when the sales become less because of its unacceptability. Supervised learning deals with Regression and Classification. Classification is being implemented in different fields Ex: Medical diagnosis and this model has gained a moment to understand the change in behaviour of the consumer, so this helps business works on the data.

Every customer invests a lot to buy a new car, its fair for him to know whether the car is worth or not for him to buy by analysing its features. If a customer buys any car, mainly that car should have more safety level and high comfort level. I am searching for a dataset which contains features of a car which decide whether a car is worth or not.

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

Problem Statement

Buying a car by investing lot of money on it and there is a chance that the customer may feel that the car he bought is not worth for its price because of its low safety level or low comfort level. Due to low safety level there may be a chance of accidents. If a customer wants to buy any particular car it would be better if he get to know whether the car is worth for him to buy or not by knowing the features of that car.

Features and Description:

Sales: Level of Sales

Maintenance: Level of maintenance

Doors: Number of doors for the car

Persons: Capacity of the car (Number of persons in a car)

lug_boot: Size of the luggage boot

Safety: Safety level of customers

Target: This is our target variable

By considering the above features Safety level is very important feature for a customer to buy a car because their nothing more important for a person more than his life. Capacity of the car and size of luggage boot is also much more important features. I think that maintenance will not impact that much. Level of sales also helpful since it describes the sales of the car.

Datasets and Inputs

The Car Evaluation Database contains examples with the structural information removed, i.e., directly relates CAR to the six input attributes: Sales, Maintenance, Doors, Persons, lug_boot, safety. I collected this data set from https://archive.ics.uci.edu to know what features are required for a car and to predict whether the car is worthy(acceptable) with that features.

Sales: Level of Sales

Maintenance: Level of maintenance

Doors: Number of doors for the car

Persons: Capacity of the car (Number of persons in a car)

lug_boot: Size of the luggage boot

Safety: Safety level of customers

Target: This is our target variable

Total number of records are 1728 and 7 columns. And the number of cars that are acceptable(worthy) are 518 and number of cars that are unacceptable(unworthy) are 1210. Here Target is our target variable which tells whether it is worth or not.

I took this data set from https://archive.ics.uci.edu and made little changes to the data set.

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

Solution statement

Buying a car is not easy for everyone since more money is needed. Once if we buy a car and later if we feel its unsafe and comfort level is less it is not easy to replace it or sell it, since the customer gets loss. So, it would be better for the customer to know its features first and know whether it is worthy or not and then buy it. So, I am using classification model of supervised learning by passing the current employee data to the model and predicting whether the car is worthy or not. There are no null values. And I will do hot encoding for categorical column and there are exactly 10 features in the dataset. I used Label encoder and converted the categorical data into numerical data. I will perform ensemble learning model to understand why the car is unworthy and will finalize only one model base on the fscore. If the model is performing better we will perform tuning and the main theme of the model is to find whether the car is worth for buying or not.

Benchmark model

Logistic regression is used as benchmark model. FBeta score of benchmark model is reference and other model will be judge to perform better if their fbeta score will be greater than Logistic regression model. Accuracy, f-beta score and confusion matrix and will try to get better results in the ensemble learning models.

Evaluation Metrics

Accuracy in classification problems is the number of correct predictions made by the model over all kinds predictions made.

$$\text{Accuracy} = (TP+TN)/(TP+TN+FP+FN)$$

In the Numerator, are our correct predictions (True positives and True Negatives) (Marked as red in the fig above) and in the denominator, are the kind of all predictions made by the algorithm (Right as well as wrong ones).

Precision is a measure that tells us what proportion of patients that we diagnosed as having cancer, actually had cancer. The predicted positives (People predicted as cancerous are TP and FP) and the people having a cancer are TP.

$$\text{Precision} = TP/(TP+FP)$$

Recall is a measure that tells us what proportion of patients that actually had cancer was diagnosed by the algorithm as having cancer. The actual positives (People having cancer are TP and FN) and the people diagnosed by the model having a cancer are TP. (Note: FN is included because the Person actually had a cancer even though the model predicted otherwise).

$$\text{Recall} = TP/(TP+FN)$$

Fβ score – measures the effectiveness of retrieval with respect to a user who attaches beta times as much importance to recall as precision.

$$F\beta=(1+\beta 2)\cdot \text{precision}\cdot \text{recall}/(\beta 2\cdot \text{precision})+\text{recall}$$

When beta=0.5 more emphasis is placed on precision. This is called f 0.5 score.

Project Design

First of all I will load the csv into a data frame with the help of pandas library and after loading it with the help of NumPy library I will calculate the number of car records that are accepted and number of records that are unaccepted, Since there is no null values in the data set there is no need of finding null values in it. And then I will encode the categorical data into numerical data. After that I will plot a heat map of all the features and will come to know how the features are correlating with each other with the help of mathplotlib.pyplot library. As all the categorical column has been converted to numerical column and all numerical column has been scaled. In the below section will split data (both features and label) into training and test sets. 80% will be used for training and 20% will be used for test. And classification model is evaluated on the basis of accuracy, precision, and recall. Logistic regression is used as benchmark model. FBeta score of benchmark model is reference and other model will be judge to perform better if their fbeta score will be greater than Logistic regression model. And with the help of ensemble learning methods will find the accuracy scores of the model. And then I varied training-to-testing data and observed the results. The results showed that as the proportionality of training data is increased, the classifier accuracy will also increase in both the cases.

References

- http://scikit-learn.org/stable/modules/ensemble.html
- https://archive.ics.uci.edu/ml/datasets/Car+Evaluation
- https://www.dataquest.io/blog/learning-curves-machine-learning/