Machine Learning Nano Degree

Capstone Project Proposal

17 March 2018

Ilyas Ahmed Mohammed

Udacity MLND

Narrative: Uncle Sam owns 'Men Junkyard' in Farland which is a very established family business. People come in and sell their cars to Uncle Sam who disassembles them and sells the components as scrap making enough money to live, laugh and enjoy. Life is good for Uncle Sam and he has no regrets expect for his little son, Smarty who left to CA two years ago to pursue further studies in Machine Learning. Uncle Sam did not approve of his choice because he didn't understand the work and he wished that Smarty would take up the family business and work on some 'manly' things rather than wasting time on computers and working with numbers.

Smarty visits his family during the Christmas holidays. Upon insistence from his mother, Smarty visits the junkyard and spends one day there managing the business while Uncle Sam is running errands for Christmas celebration. Smarty notices that majority of cars brought in by people were junk and of no use. However, some of them were in acceptable condition and he thought that with some minor repairs and modifications, those cars can be reused and sold to prospective buyers. This sparked an idea in Smarty's mind that if he could help identify which cars are in acceptable conditions, then the cars just needs to be repaired and sold, thus saving the cost of disassembling the car and the logistic cost of shipping different scrap parts for selling. Smarty wanted to help his old man and also prove to him that he can make a significant contribution to the family business and help improve its revenue. He then sets off to collect data and making a model to accurately predict the acceptability of a car to see if it can be reused/sold.

Domain Background: Car Junk sales are quite common in today's world throughout the country and one common theme in all these transaction is the ease with which a car is considered as 'junk' simply because it is not 'running'. This is mainly due to lack to knowledge of automotive among car sellers and junkyard owners. An attempt is made in this project to create a model which can predict a car acceptability rate and see if the car be reused/sold with some minor repairs. Though the primary aim of the project is to help junkyard owners identify acceptable cars and improve their revenue by selling the cars which in turn helps in creating less waste and scrap, this kind of analysis can also be used by automotive manufacturers to provide optimum warranty to customers and help them in educating about the car's performance and warranty. Working in the automotive industry for over 3 years, one of my motivation is to understand the reliability of a car and present it to the general public to save them from losing money on cars which can be easily fixed. This project is a small attempt in that direction.

Data Source: UCI Machine Learning Repository

Applications of this Data Set for similar kind of problem: Making efficient learning algorithms with exponentially many features

Problem Statement: Given a set of variables about a car's condition and some basic information about its price and comfort, how can we predict the car's acceptability rate and decide if the cars needs to be sold as junk **(or)** if the car is in acceptable condition and can be reused with some new/modified parts **(or)** if the car is in good condition and can be reused with minor repairs **(or)** if the car is in very good condition and can be sold to a pre-owned dealer or to potential buyers. This is a supervised learning classification multi-class classification problem. The input variables of the data set are some characteristics of a car such as price (buying price and maintenance price) and comfort(number of doors, estimated safety, size of luggage boot and number of person which the car can accommodate). These variables are readily available in the dataset and using these input features, the overall acceptability rate of a car needs to be predicted (CAR – Car Acceptability Rate) which is the target variable. The Target variable can be one of the four classes – unacceptable condition ('unacc'), acceptable condition ('acc'), good condition ('good') and very good condition('vgood'). A careful study of the data reveals that the majority of the target label present in the data is 'unacc' (~70%) and the next highest label is 'acc'(~22%). The remaining two labels – 'good' and 'vgood' have same distribution (~3 -4%). This shows that the data is quite skewed towards 'unacc' label.

Data Sets and Inputs: The Data is collected from UCI Machine Learning repository. It has 1728 Instances and 6 variables which are defined below:

1. Buying_price: The price of buying the car. Attributes: v-high, high, med, low
2. Maint_price: The price of the maintenance of the car. Attributes: v-high, high, med, low
3. doors: Number of doors in the car. Attributes: 2, 3, 4, 5-more
4. persons: Capacity of the car in terms of persons it can carry. Attributes: 2, 4, more
5. lug_boot: Size of the luggage boot. Attributes: small, med, big
6. safety: estimated safety of the car. Attributes: low, med, high

The dataset does not have any missing values and the entire dataset is categorical in nature and very simple which is done purposely keeping in mind the limited knowledge of the general public of all the attributes and to make it easy for them to use the model (Uncle Sam in the narrative)

Solution Statement: The model predicts the Car Acceptability Rate (CAR) of a car given its price (buying and maintenance) and its comfort characteristics(#of doors, #of persons it can carry, size of the luggage boot and the estimated safety). The target variable is CAR which has four attributes: unacc, acc, good, v-good.

For the purpose of this model, these four attributes can be defined as follows:

- unacc: The car is in unacceptable condition and it's only use is to scrap and use the spare parts
- acc: The condition of the car is acceptable and it can potentially be reused by replacing some parts
- good: The car is in good condition and can be reused with minor repairs/modifications which can be done inhouse
- vgood: The car is in very good condition and requires some regular maintenance after which it can reused or sold to a pre-owned dealer or to other prospective buyers

Benchmark model: A simple/naïve model is to predict everything which comes in to junkyard as unacceptable (unacc) which is what Uncle Sam has been doing. This can be taken as a Benchmark model and our objective is to perform better than the benchmark model in identifying cars which are in good condition and save money on disassembling the car thus increasing the revenue of the junkyard.

Evaluation Metrics: It is important that the model makes good predictions which are better than the Benchmark model. As mentioned previously, the data is skewed with over 70% labels being 'unacc' and the next three labels – 'acc', 'good' and 'vgood' occupy the remaining 30%. For this reason, accuracy might not be a good evaluation metric for this dataset since the accuracy of predicting 'unacc' will be over 70% even if everything is predicted as 'unacc' and this can be quite misleading.

It is important that the model makes accurate predictions i.e. for a car in unacceptable condition, the model should predict unacceptable otherwise defective cars will be reused/sold and this will cause serious reputation to the business and can easily back-fire. Precision and Recall are important evaluation metrics for this model and a combination of these two metrics- F-Score will be used to evaluate the overall performance of the model. Since this a multi-class classification problem, a confusion matrix will be developed and individual precision and recall for each class is calculated and then the average precision and average recall is calculated using which the overall F-score is calculated.

Project Design: At first the data is loaded in Data Frame and Exploratory Data Analysis is performed. The target variable is removed and added to a separate table. Then Data Pre-processing is done by one-hot encoding all the variables since all the variables are categorical. Due to the complete categorical nature of the data set, no logarithmic scaling or normalization techniques are required. Since the data set target labels are skewed, care needs to be taken to appropriately divide the training and test data sets. Since the majority of the target label is 'unacc', it could easily happen that all the instances in the training set could be of this label which would lead to severe underfitting of the data and the model will not perform better on test set. To prevent this problem, K-Fold Cross validation technique will be implemented when dividing the data set for training and testing.

Two or three classifiers are first tested and evaluated using the evaluation metrics stated above out of which one classifier is selected based on its performance. Then this classifier is optimized by using Grid Search technique and finding the optimum values of parameters of the classifier. Before optimizing, feature importance can also be performed to eliminate unimportant features but the dataset already has 6 features and removing some features might result in Overfitting. Finally, applicability and further enhancements to the model for improved will be recommended.