

NAME OF THE PROJECT

CAR PRICE PREDICTION

Submitted by:

V TARAK RAM SAI

**ACKNOWLEDGMENT**

This includes mentioning of all the references, research papers, data sources, professionals and other resources that helped you and guided you in completion of the project.

**INTRODUCTION**

* Business Problem Framing

This is regarding a client who works with small traders to sell used cars. The focus of this project is developing machine learning models that can accurately predict the price of a used car based on its features.

* Conceptual Background of the Domain Problem

Deciding whether a used car is worth the posted price when you see listings online can be difficult. Several factors, including mileage, make, model, year, etc. can influence the actual worth of a car. From the perspective of a seller, it is also a dilemma to price a used car appropriately. Based on existing data, the aim is to use machine learning algorithms to develop models for predicting used car prices.

* Review of Literature

In this report, we investigate the application of supervised machine learning techniques to predict the price of used cars. The predictions are based on historical data collected from online websites. In this paper, the price evaluation model based on big data analysis is proposed, which takes advantage of widely circulated vehicle data and a large number of vehicle transaction data to analyse the price data for each type of vehicles by using various Machine Learning models.

* Motivation for the Problem Undertaken

The purpose of this project is to determine the features of cars that are important to predict the cost. It aims to establish a second-hand car price evaluation model to get the price that best matches the car by observing the costs in different websites that sell used cars.

**Analytical Problem Framing**

* Mathematical/ Analytical Modelling of the Problem

Any machine learning model should follow the below steps while dealing a business problem. They are:

**i.)** **Business Understanding:** The first step is to comprehend the research’s background, the problem description, and how the proposed project will achieve the goals.

**ii.**) **Data Understanding:** The second stage requires collection of data listed in the project resources. This involves in determining the data requirements and exploring key data attributes.

**iii.**) **Data Preparation:** The third stage involves data cleaning and should the handle the missing values in the data.

**iv.**) **Modelling:** This involves determining the modelling technique and testing the design.

**v.**) **Evaluation:** Here, we should evaluate the achieved results and should determine the performance of the model with best accuracy.

**vi.**) **Deployment:** The last stage is implementation of the model.

* Data Sources and its format

**full\_name:** Brand and model of the Car.

**selling\_price:** The price of the used car.

**year:** Year of the car that is bought.

**seller\_type:** Seller type of the car(Individual or Dealer)

**km\_driven:** Total kilometres that car has travelled.

**owner\_type:** It describes the no. of people who used the car.

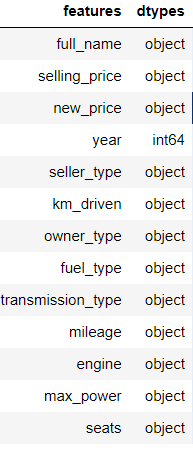
**fuel\_type:** Type of fuel that car runs by.

**mileage:** Mileage of the car.

**engine:** CC of car’s engine.

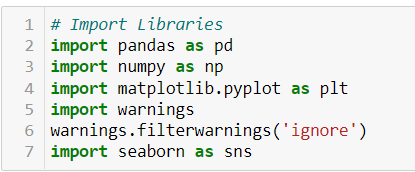
**max\_power:** Maximum power of the car.

**seats:** Total seats in the car.



* Data Pre-processing

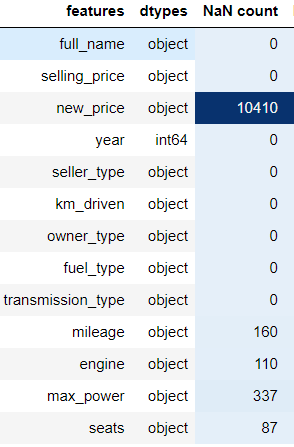
It entails converting raw data into comprehensible format that a machine learning model can understand. The data pre-processing involves data cleaning which involves handling missing values, transformation of data i.e. normalizing the data and data reduction which involves only required features.





**Data Cleaning**

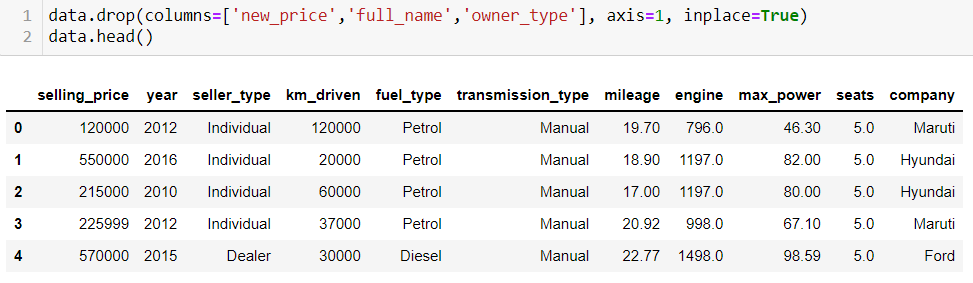
The first step of pre-processing is data cleaning by checking and eliminating any missing values because they affect the accuracy of the model. This is achieved by either filling the missing values with a mean or mode function or by dropping all the missing values. In this case, there are missing values in new\_price, mileage, engine, max\_power and seats columns.



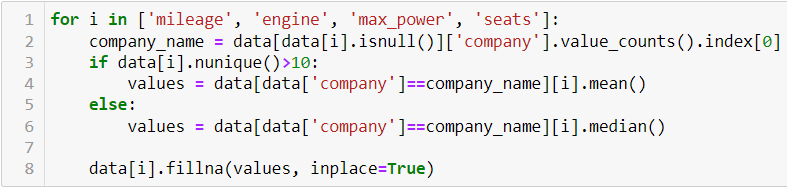
**Data Reduction**

The next step of data processing is data reduction. This is used to remove duplicate features present in the data i.e. unwanted features for prediction.

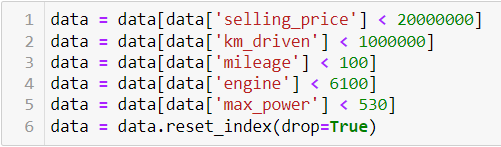
Dropping "new\_price" column which has more the 50 percent missing values, "full\_name" because I created company column for better analysis and "owner\_type" for having only one value which doesn't provide any information.



### Handling Missing Values



**Handling Outliers**



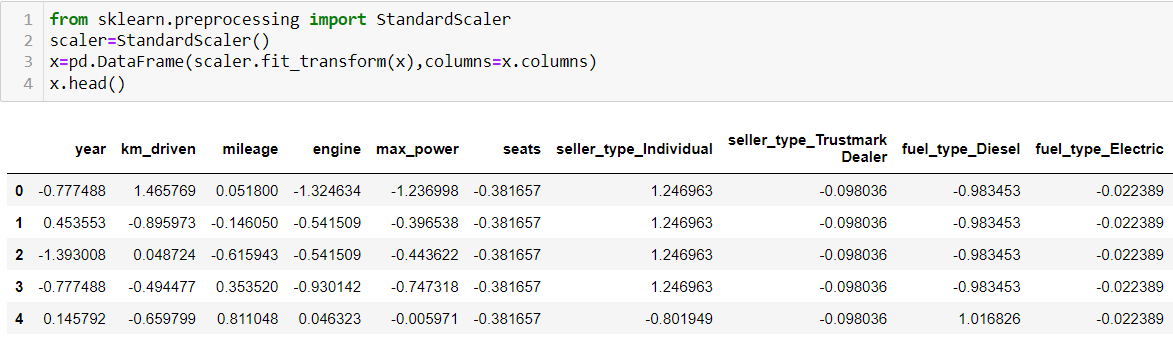
**Encoding**

Encoding categorical data is a process of converting categorical data into integer format so that the data with converted categorical values can be provided to the different models.



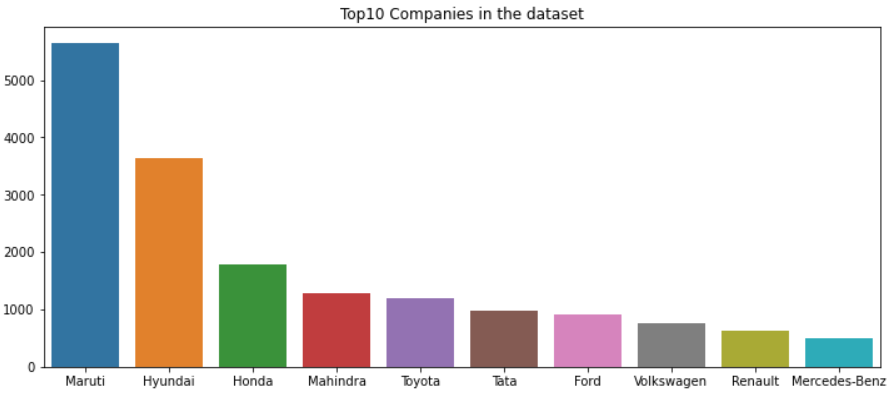
**Scaling**

Feature scaling is **a method used to normalize the range of independent variables or features of data**. To convert data into a distribution with a mean of 0 and standard deviation of 1, we will use a standard scalar.

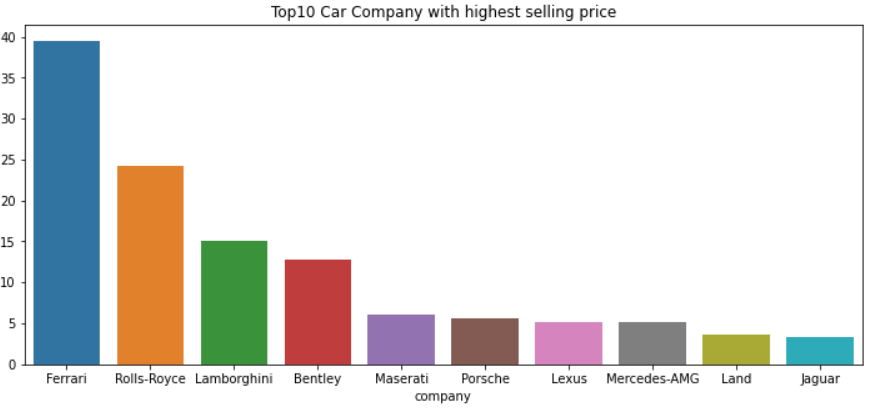


* Data Inputs- Logic- Output Relationships

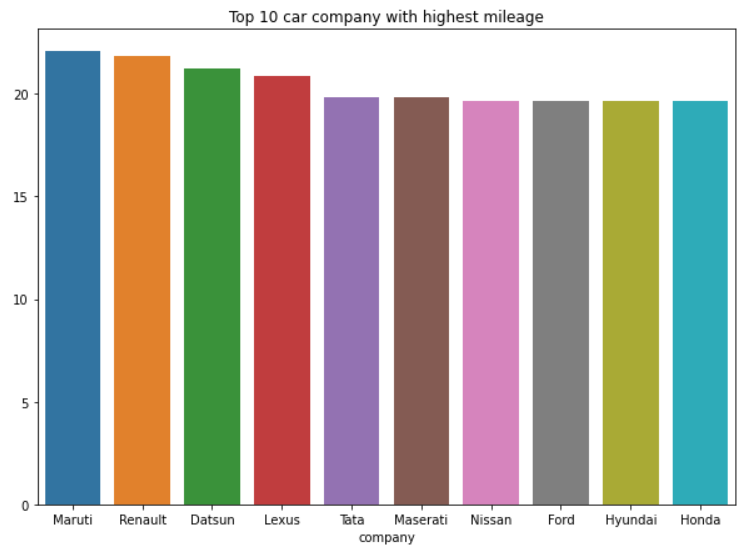
The Relation between all the features in the dataset are determined in the below graphs as follows



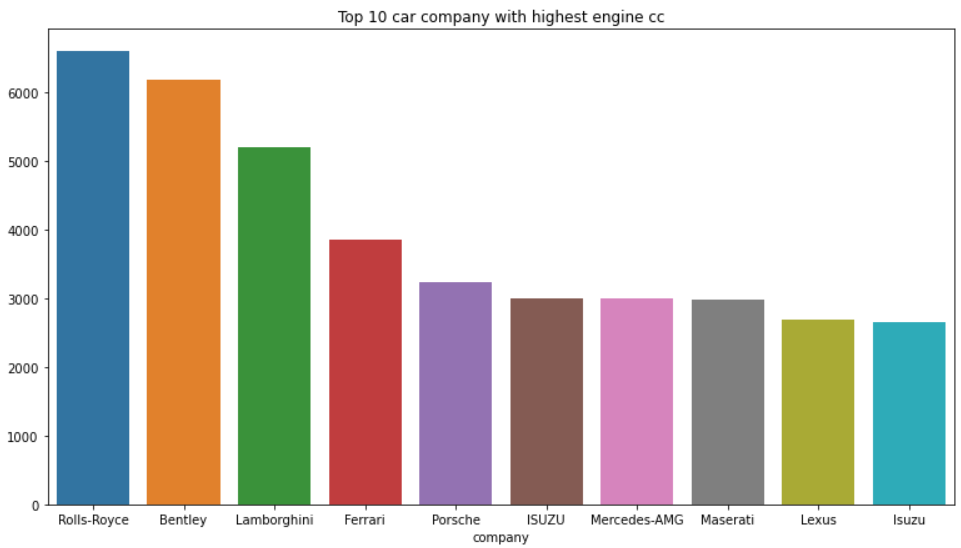
These are the Top 10 companies present in the dataset.



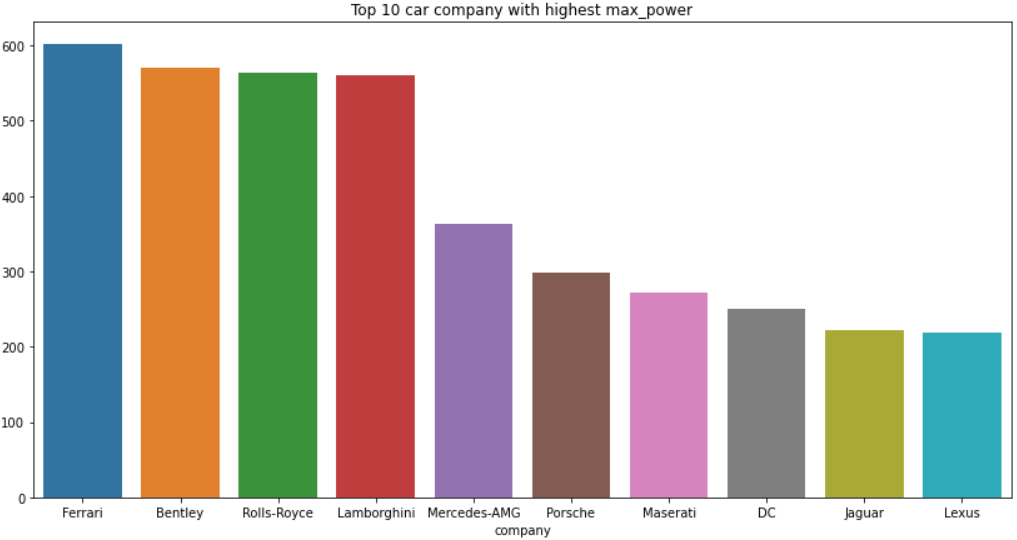
These are the Top 10 companies with highest selling price.



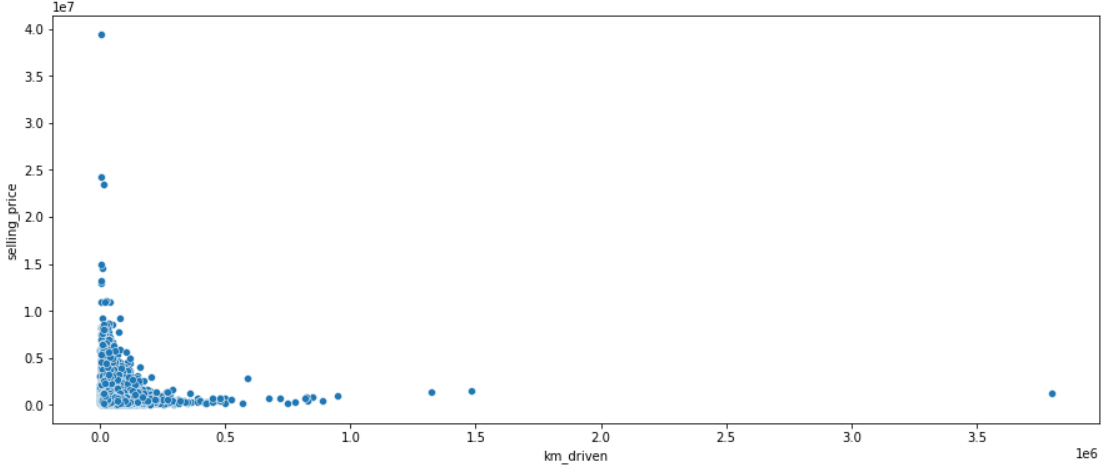
These are Top 10 car companies with highest mileage.



These are Top 10 car companies with highest engine cc.



These are Top 10 car companies with highest maximum power.



### If kilometres driven is more, then the selling price of the car will be less.

### 

### There is no significant relation between mileage and selling price of the car. In some cases, higher the mileage lower is the selling price.

### 

### If the engine's cc is more, then the selling price of the car is also more.

### 

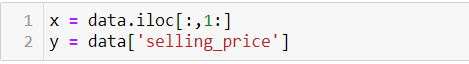
### If the power of the engine is more, then the selling price of the car is also more.

* Hardware and Software Requirements and Tools Used

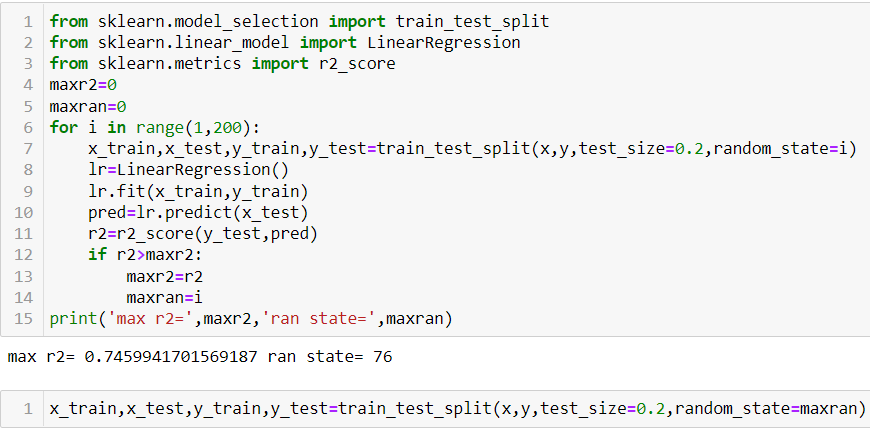
The hardware requirements for the project includes a laptop with at least 4GB RAM. This project uses a Jupyter Notebook as a code editor. The Machine Learning models are implemented using python version 3.7 with libraries like numpy, pandas, matplotlib, seaborn and sklearn.

**Model/s Development and Evaluation**

The independent variables are declared in x and the dependent variable i.e. ‘selling\_price’ is declared in y as follows-



The below code is done for choosing the Random state variable. We should do testing by using any of the four classification algorithms.



**Hyper Parameter Tuning**

In machine learning, hyper parameter optimization or tuning is **the problem of choosing a set of optimal hyper parameters for a learning algorithm**. A hyper parameter is a parameter whose value is used to control the learning process. By contrast, the values of other parameters (typically node weights) are learned.

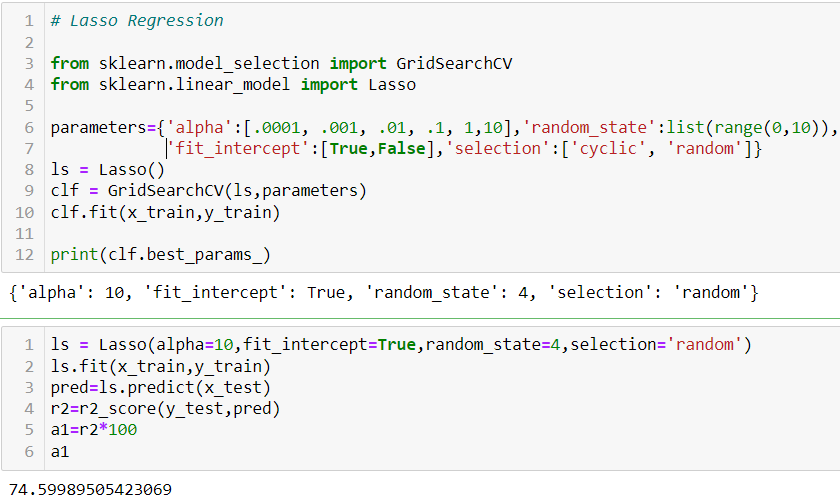
**Linear Regression**

Linear Regression attempt to model the relationship between two variables by fitting a linear equation to observed data. The other is considered to be dependent variable.



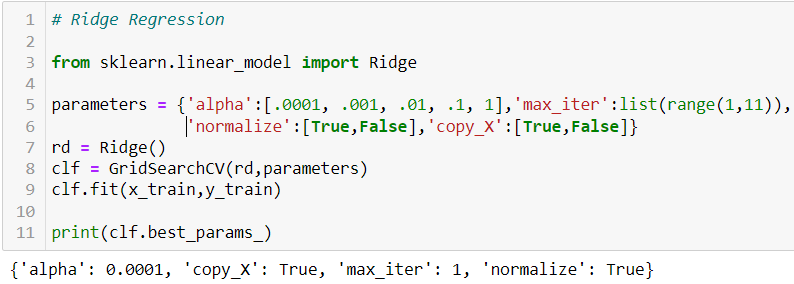
**Lasso Regression**

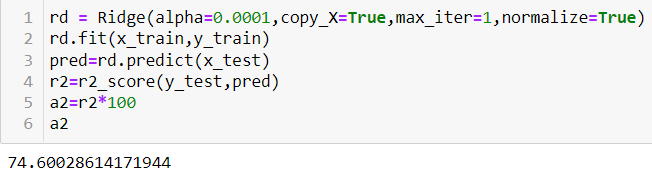
The “LASSO” stands for Least Absolute Shrinkage and Selection Operator. Lasso regression is a regularization technique. It is used over regression methods for a more accurate prediction. This model uses shrinkage. Shrinkage is where data values are shrunk towards a central point as the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This particular type of regression is well-suited for models showing high levels of multi-collinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination.



**Ridge Regression**

A Ridge regressor is basically a regularized version of Linear Regressor. The regularized term has the parameter ‘alpha’ which controls the regularization of the model i.e. helps in reducing the variance of the estimates.





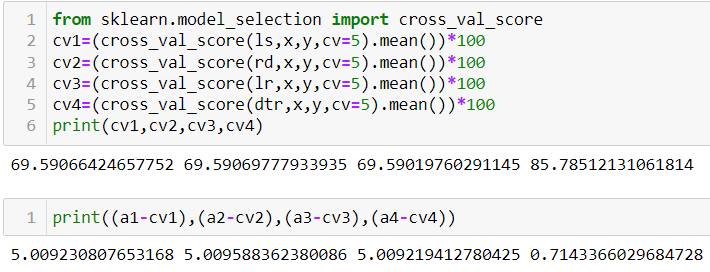
**Decision Tree Regression**

Decision tree builds regression or classification models in the form of a tree structure. It **breaks down a dataset into smaller and smaller subsets** while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes.



**Cross Validation**

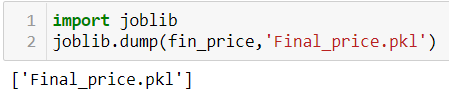
Cross Validation is a **very useful technique for assessing the effectiveness of your model**, particularly in cases where you need to mitigate overfitting.



### We can choose Decision Tree Regression as our model since its cross validation score and accuracy score are almost similar.

### 

**Saving the Model**

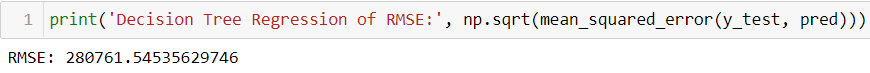


The above code will save the model.

**CONCLUSION**

* Key Findings and Conclusions of the Study

The results of our tests were quantified in terms of the R score of our predictions. R2 score is a statistical measure of how close the data are to the fitted regression line. Compared to Linear Regression, Decision-Tree based methods performed comparably well.



* Learning Outcomes of the Study in respect of Data Science

Car price prediction can be a challenging task due to the high number of attributes that should be considered for the accurate prediction. The major step in the prediction process is collection and preprocessing of the data. In this research, PHP scripts were built to normalize, standardize and clean data to avoid unnecessary noise for machine learning algorithms.

Data cleaning is one of the processes that increases prediction performance, yet insufficient for the cases of complex data sets as the one in this research. Applying single machine algorithm on the data set accuracy was about 75%. Therefore, the ensemble of multiple machine learning algorithms has been proposed and this combination of ML methods gains accuracy of 88 %. This is significant improvement compared to single machine learning method approach.

* Limitations of this work and Scope for Future Work

For better performance, we plan to judiciously design deep learning network structures, use adaptive learning rates and train on clusters of data rather than the whole dataset. To correct for overfitting in Random Forest, different selections of features and number of trees will be tested to check for change in performance.

However, the drawback of the proposed system is that it consumes much more computational resources than single machine learning algorithm.