A mathematical study of machine learning and deep learning algorithms and its effectiveness in prediction and classification of data

Devansh Sanghvi
Department Of Biotechnology
Indian Institute of Technology, Madras
Chennai, India
be19b002@smail.iitm.ac.in

***Abstract*—This document aims to study the mathematical properties of various machine learning and deep learning algorithms and demonstrate its effectiveness in using the features in the given datasets to perform classification and prediction.**

This paper will serve as a part of the requirements to be fulfilled for receiving credits in the course **EE4708: Data Analytics Laboratory.**

I. Linear Regression

***Abstract*—This paper aims to develop an understanding of the mathematical properties of linear regression and showcase its use in estimating the relationships between the various factors that have an effect on the health data obtained from the United States. Improved the linear regression model by including valuable feature selection. Significantly improved the data visualization graphs and included a visual representation of the correlation matrix of the data.**

Introduction

This paper seeks to explore one of the fundamental modelling tools used for data analysis - **Linear Regression**. Once we have acquired data with multiple variables, one important task is to understand how the variables are related. Regression is a statistical method used to determine the strength of the relationship between one or more independent variables and a dependent variable. Linear regression, one of the well-known machine learning techniques, makes the assumption that the variables have a linear relationship and accordingly gauges the relationship between the variables. Despite its simplicity, linear regression has proven itself a powerful tool for analyzing data and revealing trends in the data. Its simplicity is evident in its representation:

$$y = \beta_0 + \beta_1 X$$

In the above equation, y denotes the independent variable and X denotes the dependent variables (s) where multiple dependent variables can be succinctly expressed as a vector. The parameters of this model are β_0 and β_1 which represent the intercept/constant and the slope/scaling factor(s) associated with the dependent variable(s). For a given data point X and a given set of parameter values β_0 and β_1, we calculate the estimated value of y. The objective of linear regression is to identify the parameter values for which the sum total of errors for all data points is minimum. These values provide us a quantitative description of the relationship between the variables.
To demonstrate the effectiveness of linear regression, we use standard socioeconomic variables like income and health insurance access in a population to identify putative correlations with the incidence of cancer and mortality caused by cancer in the same population. The pertinent data has been collected from [1]. Before performing data analysis using linear regression, we clean and visualise the data.

Linear Regression

It was stated in the introduction section that despite its simplicity, linear regression has proven itself a powerful tool in data analysis. Inside its simplicity lies certain assumptions that must be mentioned before we delve into the mathematical details of linear regression.

A. Assumptions

- The first assumption is the independence of observations which means that there is no relationship between different independent variables. To be certain that this assumption is appropriate for the model we seek, one must look at the data collection process and see if it is collected without bias. Correlations between independent variables will imply redundancy and this could result in overfitting of the model
- The second assumption, which has been mentioned in the introduction, is the linear relationship between the independent and dependent variables. The assumed linearity is parametric linearity and not variable linearity. Hence, equations such as $y = \beta_0 + \beta_1 x^2 + \beta_2 x^3$ are allowed because y is linearly dependent on the parameters β_0, β_1 and β_2, and variables x^2 and x^3.
- The third assumption is that the error terms have constant variance which is also called homoskedasticity. If error terms have varying variance called as heteroskedasticity, the model will be accurate in some parts of the dataset.

B. Errors

The objective of linear regression is to identify the parameter values for which the sum total of errors for all data points is minimum. There are different ways to measure the error:

1) Mean Absolute Error (MAE) = $\frac{1}{n}\sum_{i=1}^{n}|y_{pred} - y_{actual}|$
2) Mean Squared Error (MSE) = $\frac{1}{n}\sum_{i=1}^{n}(y_{pred} - y_{actual})^2$
3) Root Mean Squared Error (RMSE) = $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_{pred} - y_{actual})^2}$
4) Mean Percent Error (MPE) = $\frac{100\%}{n}\sum_{i=1}^{n}(\frac{y_{actual}-y_{pred}}{y_{actual}})$
5) Mean Average Percent Error (MAPE) = $\frac{100\%}{n}\sum_{i=1}^{n}(\frac{y_{actual}-y_{pred}}{y_{actual}})^2$

MPE is used to check if the model's performance is symmetric because it does not take absolute values or squares of errors.

Another metric used to evaluate the model's performance is coefficient of determination R^2 which is defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_{actual} - y_{pred})^2}{\sum_{i=1}^{n}(y_{actual} - y_{mean})^2}$$

The values of R^2 lie between -1 and 1. If the R^2 of the linear regression model is closer to 1, then it is able to explain a higher proportion of the variance in y and consequently performs better in estimating y.

Does socioeconomic status determine cancer risk?

In this section, we see the correlations and relationships revealed by the linear regression models between socioeconomic status and cancer incidence, mortality. We want to see whether poor sections of the population and those without health insurance are likely to get cancer. Two important socioeconomic variables used are:

1) Incidence rate: defined as the number of cancer cases recorded in a year per 100,000 people at risk. The population is taken to be the denominator assuming that cancer is not related to age.
2) Mortality rate: defined as the number of deaths in a population in a given year per 100,000 people.

Preparing the data for analysis

Data collection is rarely perfect and since this data has been collected from thousands of counties across the United States, some pre-processing has to be done before visualising the data. In the given dataset, there are a lot of missing data entries owing to confidentiality measures. Most of these missing entries occur in the fields concerning race which are not of immediate interest to us. Hence, using the columns that interest us, we created a new dataframe and dropped the rows with NaN entries.

Visualisation

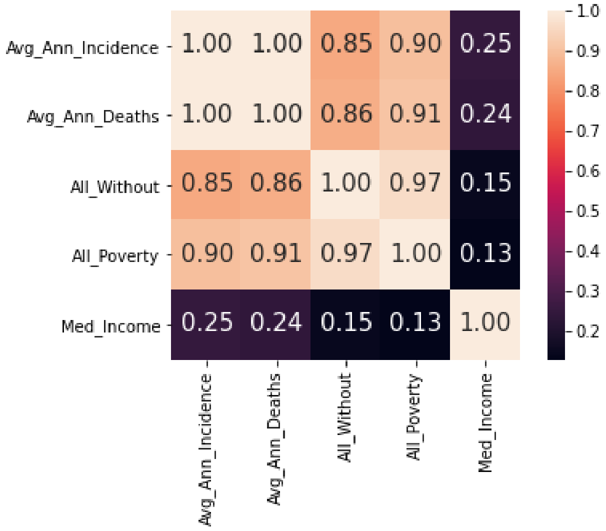


Fig. 1. Correlation Matrix visualization

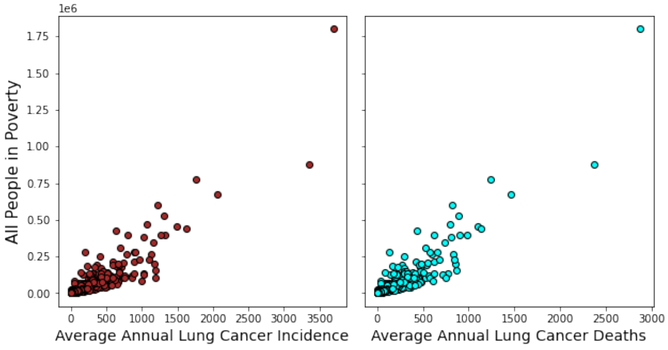


Fig. 2. Poverty numbers v/s Cancer Incidence and Deaths

First, we obtain the population of each county by adding numbers of people with and without health insurance. Poverty rate is calculated by dividing number of people in poverty by population size. The picture itself shows a positive correlation between these two variables which the model will confirm. Second, plot the cancer mortality rate against poverty rate. Again, there is a positive correlation between these two variables which the model will confirm. The plot of cancer incidence rate against median income shows clear negative correlation and so does the plot between cancer mortality rate and median income (not shown here).
The visual data shows that the poor people are likely to get cancer or die because of it. Now we have to show it with statistical proof using a linear regression model.

The linear regression model

A simple linear regression model was used with median income as X and mortality rate as y. The downward

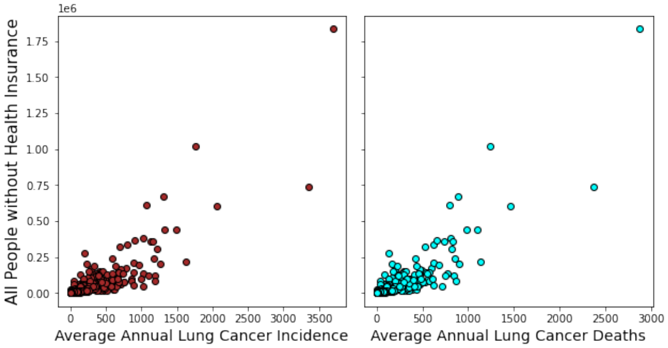


Fig. 3. People without Health Insurance v/s Cancer Incidence and Deaths

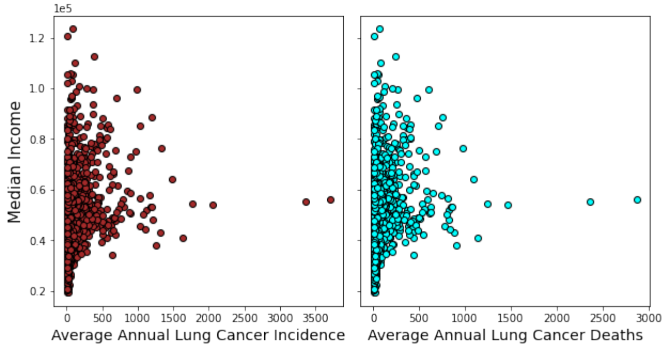


Fig. 4. Median Income v/s Cancer Incidence and Deaths

sloping lines shows a negative coefficient for X with b1 = -0.00049949. Another simple linear regression model was used with median income as X and incidence rate as y. The downward sloping lines shows a negative coefficient for X with b1 = -0.00053473. From the quantitative results obtained and the previous visualisation plots, we can see that the distribution is not really homoscedastic. As discussed, this can really hinder the performance of a linear regression model.

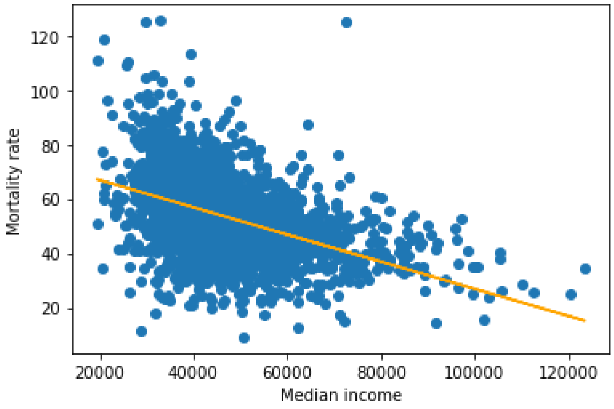


Fig. 5. Plot of median income against cancer mortality rate.

Conclusion

While the model's performance was successfully demonstrated, both visually and experimentally, that there exists a negative correlation between incidence rate and median income, and between mortality rate and median income. The results of the linear regression model are not satisfactory so the relationships in question are not, strictly speaking, linear. This can be attributed to flaws in the data collection or in the model.

After seeing the plots, I suggest a logistic regression model will serve the data better. The heteroscedastic nature of the incidence vs median income and mortality vs median income plots may also suggest the need to take into account additional features such as race or sex, but race specific incidence and death rates must be obtained first.

For the purposes of fundraising for the NGO, it can be argued that visual proof is more compelling than statistical evidence. Hence, while we can discuss the nature of the relationship between income and rates of incidence or mortality, it is clear that the mortality and incidence rates decrease with an increase in income. The figures make a compelling case for the health authorities to prioritise improving the health standard of the poorer section of the American society.

II. Logistic Regression

Abstract—**This paper aims to develop an understanding of the mathematical properties of logistic regression and showcase its application in classifying whether a passenger aboard the Titanic survived. Changes were made in the 'logistic regression model.Two plots: Feature ranking line graph and a plot with the GridSearch CV using multiple scorers were added.**

Introduction

Logistic Regression is a technique for modelling the likelihood of an event. It helps in understanding the relationship between the features and the target variables (survived or not in our case). The structure of logistic regression is very similar to the structure of linear regression. You have a set of explanatory variables (X1,X2....Xn) and our target binary variable (Y). The function behind it is more complicated than linear regression:

$$P(Y=1) = \frac{1}{1+e^{-(b+B*X)}}$$

P(Y=1) is the probability that the predicted class is 1, b is a constant that is not related to X and B represents the weights for the relationship between the X's and the Y. The logistic regression estimates the value of the constant, b and the weights using Maximum Likelihood Estimator. Upon calculating the weight and the the probability P(Y=1) for new data points and depending upon the threshold, we decide if the predicted class should be 1 or 0 (survived/ not survived in our case). For example: if the threshold to classify the new data point is 0.5 and we get P(Y=1)= 0.75, we will classify that particular data point as 1.
To demonstrate the effectiveness of logistic regression, we analyze the "Titanic" data-set provided to us. The data contains the passenger details of the 891 passengers that were aboard the Titanic on the eve of the infamous shipwreck. Passenger details included their name, sex, age, family details and their ticket details like their cabin, their passenger class, where they embarked from and their ticket prices. Another column in the data-set mentions if the passenger survived the Titanic crash or not. With the training data-set of the given passengers, we aim to predict if a given passenger would have survived the Titanic crash or not.

Logistic Regression

Logistic regression uses the maximum likelihood estimation to select the best fit line. It converts the probability into log(odds) using the logit function. To find the log(odds) value for each candidate, we project them onto the log odds value. Once we find the log (odds) value for each candidate, we'll convert the log(odds) to the probability using the sigmoid function.

$$log(\frac{p}{1-p}) = log(odds)$$

On finding the likelihood for the first line, we then rotate the log(odds) line by a bit and again calculate the likelihood. This is done using the Gradient Descent algorithm. For a particular slope, we calculate the cost of the classification prediction, and repeat the process until the best- fit log(odds) line is found. Certain assumptions are necessary for Logistic Regression:

A. Assumptions

- The dependent variable must be categorical.
- The model should have little or no multi-collinearity i.e. the independent variables should not be correlated with each other.
- The independent variables are linearly related to the log(odds).
- Logistic Regression requires quite large sample sizes.

B. Loss function in logistic regression

Mean squared error or mean squared errors cannot be used in logistic regression since the curve obtained from plotting the Mean squared error loss with respect to the logistic regression model weights is not a convex curve and therefore it is very difficult to find the global minimum. The non-convex nature of Mean Squared Error in logistic regression is because of the non collinearity introduced in the model because of the sigmoid function introduced in the model, making the relationship between the weight parameters and errors very complex. Hence the loss used by logistic regression (Binary Cross-Entropy Loss/ Log Loss):

$$\frac{-1}{N}\sum_{i=1}^{n} y_i \cdot log(p(y_i)) + (1-y_i) \cdot log(1-p(y_i))$$

C. Feature selection in logistic regression

Every raw data-set contains a lot of redundant features, that may impact the performance of the model. Feature selection is a component of feature engineering that involves the removal of irrelevant features and picks the best set of features for the model. In other words, it helps reduce the dimensionality of the data. There are various techniques used for feature selection, but one common method is the Recursive feature elimination: It is used to select features by recursively considering smaller and smaller sets of features. We do this by eliminating the most useless features in our data in each iteration until we reach the desired amount of features.

D. Review of model evaluation procedures

To choose between different machine learning models, we need to evaluate the models using the right procedure. The simplest approach is to train and test on the same data, but this method results in overfitting the data. An alternative approach will be to split the data into train and test data. This is because testing accuracy is a better estimate of the model than the training accuracy. There are a couple of problems with train test split such as:

- It provides a high variance estimate
- Testing accuracy can change a lot depending on which data points are chosen in the test data-set.

While splitting the data into train/test data has its own shortcomings, they are still better than testing on the same data as training data and hence are used more often.

Preprocessing and exploratory analysis on the data

In this section, we will first preprocess the data to extract the important features in the data provided, then we will visualize the data to understand the effects of certain features and modify the data-set to optimize the results from the logistic regression model.

Preprocessing the data

The initial training data-set has 891 samples and 12 features/ passenger details, namely: PassengerId, Survived, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin and Embarked. Similarly, the testing data-set has 418 samples. Preprocessing the data-set includes (Note that the changes are made in both the training and testing data-sets):

1) Missing value assessment: Some features have a percentage of missing values and the first step is to decide what to do for these features. Age, Embarked both have a small percentage of NA values (19.87 and 0.22 respectively). We replace the missing age values by the median age of the known ages, 28.00 and impute the missing Embarked values by the port on which the most people boarded, S. Cabin has 77.10 percent of its values missing, and hence it is appropriate to ignore the feature altogether.
2) Additional Variables: Both Parch and SibSp are related to if the passenger is travelling with a family or alone, and the two features can be combined into one feature: if he was travelling alone or not. For the subjective features: gender, passenger class and embarked, we create categorical features to represent them numerically.

Visualization

After preprocessing the data, we visualize the data, to better understand the correlation between different features:

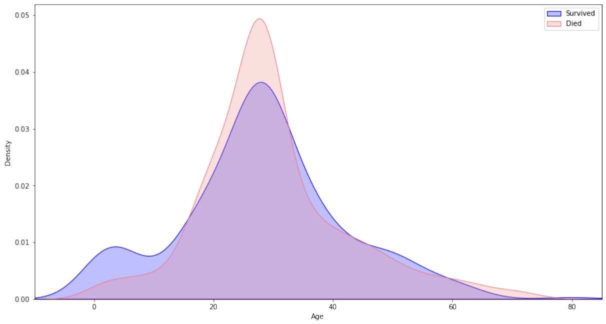


Fig. 6. Density Plot of Age for Surviving Population and Deceased Population

1) Exploration of Age: The two graphs show very similar distributions with respect to the age, but one noticeable difference is that a larger fraction of the survived population are children. It is therefore practical to add a feature to indicate if the passenger was a child or no.

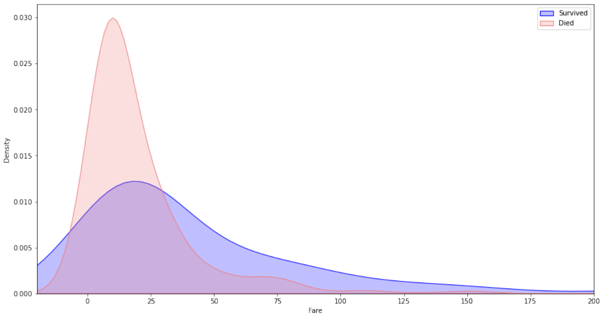


Fig. 7. Density Plot of Fare for Surviving Population and Deceased Population

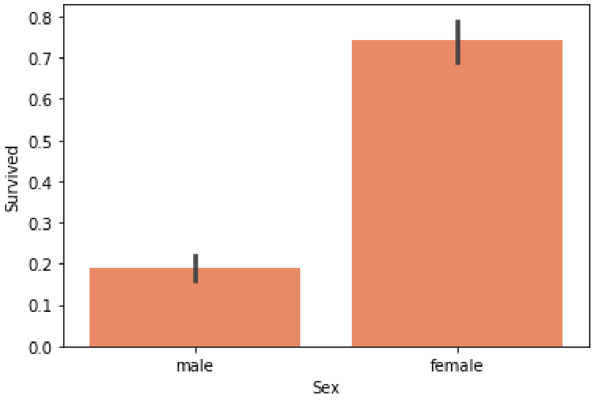


Fig. 8. Bar Plot of the fraction of the two gender populations that survived the crash

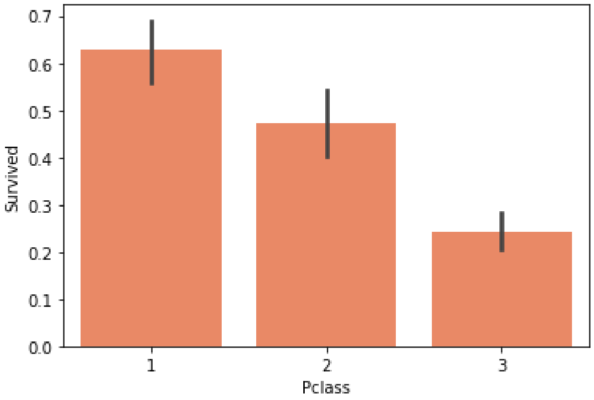


Fig. 9. Bar Plot of the fraction of passengers in each class that survived the crash

2) Exploration of Fare: Passengers who paid less are less likely to survive and hence this could be an important indicator for our predictions.
3) Exploration of Gender: It can be clearly seen that females were preferred in the rescue procedure.
4) Exploration of Pclass: Fare and Pclass reiterate the same thing: the amount of money paid for a ticket is a strong indication of if the passenger survived.
5) Exploration of Embarked: People who boarded the ship

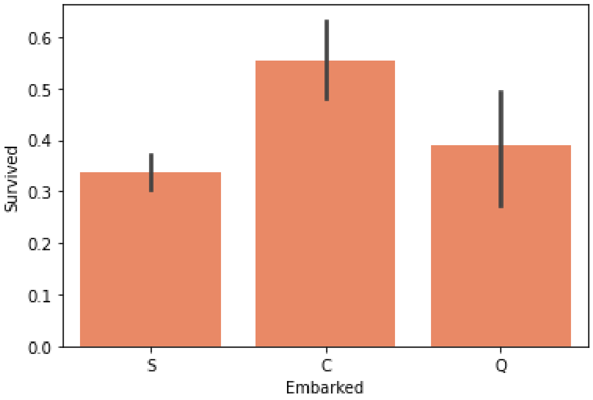


Fig. 10. Bar Plot comparing the fraction of passengers that survived with their respective ports they embarked from

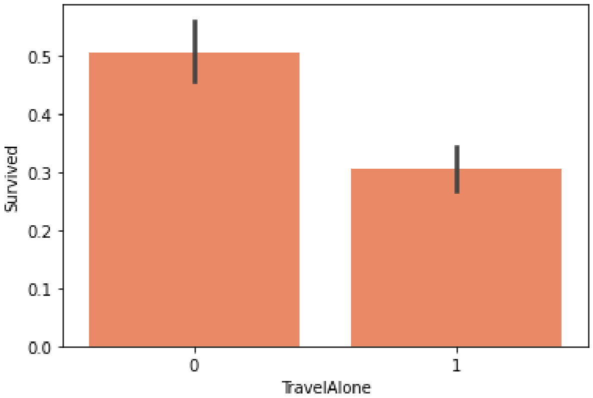


Fig. 11. Bar Plot of the fraction of passengers that survived while travelling alone or with family

in Cherbourg, France, appear to have the highest survival rate, and people who boarded in Southampton were less likely to survive than the people who boarded in Queenstown.

6) Exploration of Travelling Alone: People without a family were more likely to die in the crash.

The Logistic Regression Model

After preprocessing the data, we move to the heart of the problem: the prediction of whether an unseen passenger (not in the training data-set) survived the crash. Here are the steps for the same:

- Feature Selection: Before applying a logistic regression model to our data, we will eliminate the features that are least important using recursive feature elimination. The 8 remaining features are: Age,TravelAlone, Passenger classes, Embarked, Sex, IsMinor. Therefore, it is best to not include Fare in our model prediction. The correlation matrix for the same can be seen in image Fig.7.
- Logistic Regression model and results: After selecting the features, we run the logistic regression model using the simple test/ train model and evaluate our results using three parameters, which turn out to be the following:
  1) accuracy: 0.782
  2) logloss: 0504
  3) auc: 0.839

  Using the same model, and including Fare as a feature in our model we get the following results:
  1) accuracy: 0.796
  2) logloss: 0.455
  3) auc: 0.849

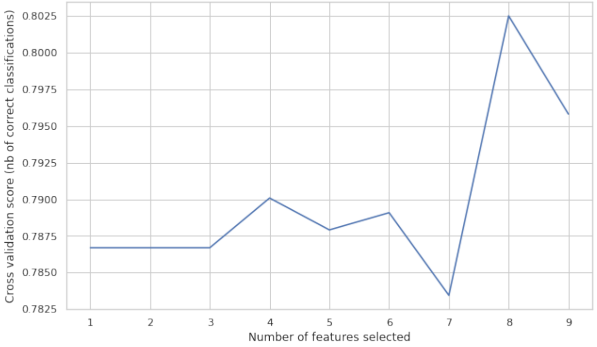


Fig. 12. Feature ranking for the logistic regression model

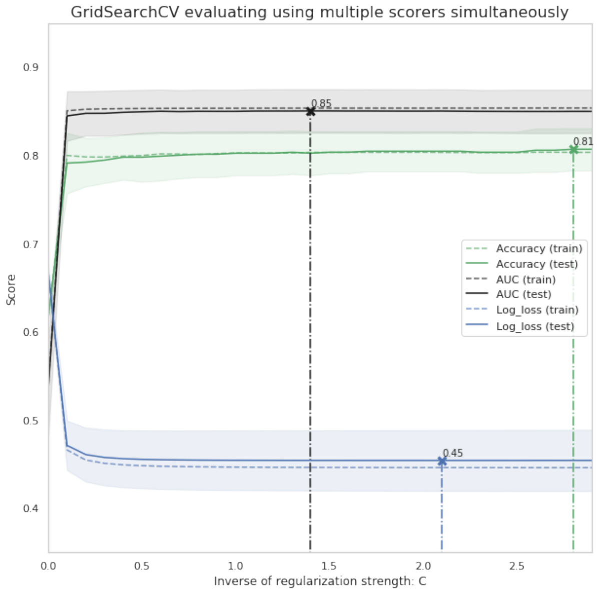


Fig. 13. Gridsearch CV evaluating multiple scorers simultaneously

Conclusion

The logistic regression model was succesfully demonstrated, both visually and numerically, that there exists a correlation between gender and survival, and between the passenger class and survival. The addition of the L1 regularized term and the use of SAGA solver would result in a small increase in the accuracy of the model.

III. Naive Bayes’ Classifiers

***Abstract*—This paper aims to develop an understanding of the mathematical properties of Naive Bayes Classifier and showcase its application in classifying the income of people from various walks of life. Replaced the previous visualization plots with more meaningful and colourful plots.**

Introduction

Naive Bayes is a very simple, yet effective and commonly used machine learning classifier. It is a probabilistic classifier that uses Maximum A Posteriori decision rule to take a decision in the Bayesian setting. Using the features x_0 through x_n and classes c_0 through c_n, we calculate the probability of the features occurring in each class, and to return the most likely class. Hence, we want to calculate P(c_i | x_0,.....x_n). The required probability can be calculated using the provided data and applying the formula (**Bayes Rule**):

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

According to the equation, probability of the event A (class C_i in our case) given an event B (for certain values of features) is equal to probability of B given A, multiplied by P(A), divide by P(B). Since, P(B) acts as a normalization, we can ignore that term and instead state that P(c_i| x_0,.. x_n) is proportional to P(x_0,.... x_n | c_i). To simplify this computation, we assume that x_0 through x_n are **conditionally independent** given c_i and our final representation for the Naive Bayes Classifier is:

$$P(c_i|x_0, ...x_n) \propto P(x_0, ...x_n|c_i)P(c_i)$$

$$\propto P(c_i)\prod_{j=1}^{n} P(x_j|c_i)$$

To demonstrate the effectiveness of Naive Bayes Classifiers, we analyze the ”adult.csv” file that has been provided to us. The data-set contains the data of **32561 adults** across the world with personal details like: gender, level of education, their occupation, marital status amongst others. With the training data-set of the adults, we aim to **predict if the income of a particular adult is more or less than 50K**.

Naive Bayes Classifier

Naive Bayes Classifier uses the Bayes Rule to determine the probability of the occurrence of each class using the feature values given. Once we find the respective probabilities, we define regions in n-dimensions for each class. A point in a particular region will be classified as class c_i if the probability of class c_i is the most in that region. Next, we list down the assumptions, pros and cons of Naive Bayes Classifier:

A. Assumptions

- The fundamental Naive Bayes algorithm makes an assumption that each feature makes an independent contribution to the outcome.
- For numerical features, normal distribution is assumed.
- The fundamental Naive Bayes algorithm makes an assumption that each feature makes an equal contribution to the outcome.

B. Advantages of Naive Bayes

- It is easy and fast to predict the class of the test data set. It also performs well in multi class prediction.
- When the assumption of independence holds, a Naive Bayes classifier performs better when compared to other models like logistic regression.
- Less training data is required than other classification algorithms.

C. Disadvantages of Naive Bayes

- If the categorical variable has a category (in test data set), which was not observed in the training data set, then the model assigns a 0 (zero) probability and will be unable to make a prediction. This is often known as “Zero Frequency”.
- Naive Bayes is also known as a bad estimator, so the probability outputs from Naive Bayes’ function are not to be taken too seriously.
- Another shortcoming of Naive Bayes is the assumption that each feature is independent of each other. It is almost impossible to get features which are completely independent in the real world.

D. Evaluation Parameters in Naive Bayes Classifiers

There are multiple Bayes Classifiers that are used, and an evaluation metric is required to compare the classifiers. Some metrics that can be used to compare are:

- Confusion matrix: Four values are present in the confusion matrix.a_11 represents the true positive (TP) and a_22 represents the true negatives (TN) value. These are the respective fractions of correct predictions when compared to the actual values i.e the prediction was true when the actual classification was true, in the case of true positives. a_12 represents false positives (FP), i.e the number of times the actual value was false, but the value predicted was true. a_21 represents true negatives (TN), i.e the number of times the actual value was true, but the value predicted was false.

Preprocessing the data

The initial data-set has 32561 samples and 15 features, namely: Age, workclass, fnlwgt, education, education.num, marital status, occupation, relationship, sex, race, capital.gain, capital.loss, hours.per.week, native.country and income. Preprocessing the data-set includes (Note that the changes are made in both the training and testing data-sets):

1) Missing value assessment: Some features have a percentage of missing values. While none of the features have missing values, the features workclass, occupation and native.country have ”?” values. We replace the workclass and native.country ”?” values with the modes of these features respectively. We replace the occupation ”?” values with the ’Adm-clerical’ occupation.
2) Additional Variables: We do not require the ’education’ feature since we know the respective education.num. Similarly ’fnlwgt’ feature is unneccesary.

Visualization

After preprocessing the data, we visualize the data, to better understand the correlation between different features:

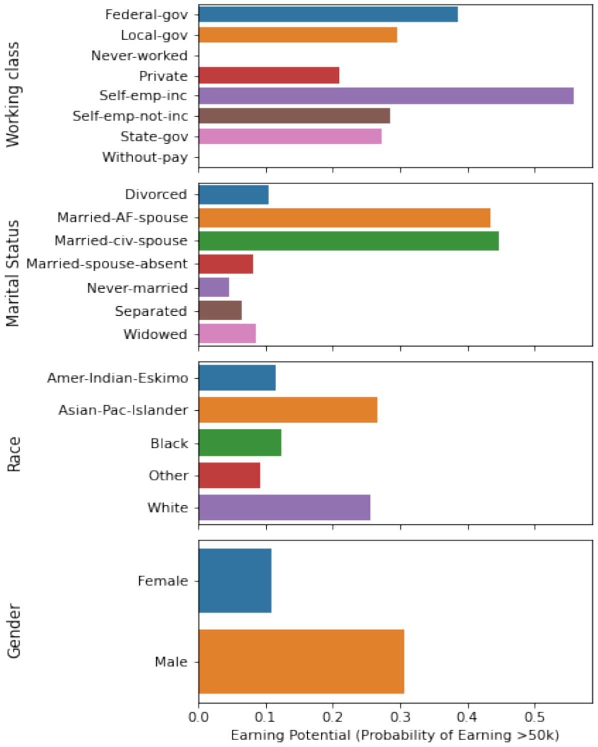


Fig. 14. Box Plot of the data

The Naive Bayes Model

After preprocessing the data, we move to the heart of the problem: the prediction of whether an unseen adult (not in the training data-set) survived the crash. Here are the steps for the same:

- Feature Selection: Before applying a Bayes Classifier model to our data, we will eliminate the features that are least important using recursive feature elimination. Income levels of adults are not correlated with workclass and the native country and hence can be ignored in the final Classifier model. The correlation matrix for the same can be seen in image Fig.4.
- Naive Bayes model and results: After selecting the features, we run three different Naive Bayes models using the simple test/ train model and evaluate our results using three parameters (the values are weighted averages), which turn out to be the following:

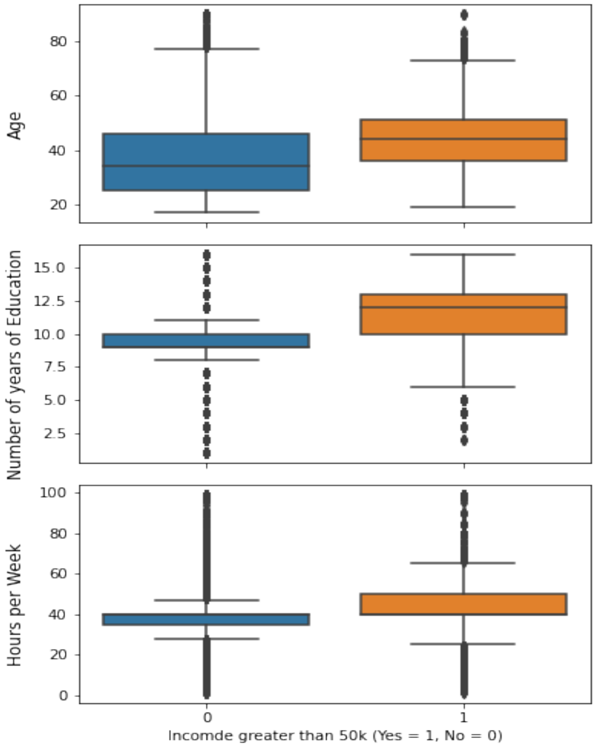


Fig. 15. A combined bar plot of the entire data

1) Gaussian Bayes:
   - Precision: 0.79
   - Recall: 0.8
   - F1: 0.77
   - Accuracy: 0.73
2) Bernoulli Bayes:
   - Precision: 0.79
   - Recall: 0.73
   - F1: 0.75
   - Accuracy: 0.80
3) Multinomial Bayes:
   - Precision: 0.81
   - Recall: 0.77
   - F1: 0.68
   - Accuracy: 0.77

Conclusion

The Naive Bayes model was successfully demonstrated, both visually and numerically. Bernoulli Bayes model proved to have the highest accuracy, while Gaussian Bayes had the smallest accuracy. Gaussian Bayes, however, had the highest F1 score. All three models had similar precision and recall.

IV. Decision Trees

***Abstract*—This paper aims to develop an understanding of the mathematical properties of decision trees and showcase its application in classifying a car based on its safety. Trees with different relevant hyperparameters (variable max depth) were made and their accuracies were compared.**

Introduction

Decision Trees is one of the most important tools used for classification and prediction. Once we receive data with some independent variables and a target variable which we want to determine for a given data point, we use statistical tools to develop a relationship between the target and the independent variables to decide on mathematical rules to assign a given data point to a class. The decision tree classifier uses a tree like structure to perform this classification.

Decision trees have applications spanning multiple fields . In general, they are constructed using an algorithmic approach that identifies ways to split the data set. There are two common attribute selection measures, namely: entropy and Gini Index. The formula and their mathematical interpretation is given below.

$$Gini = 1 - \sum_{i=1}^{n} p^2(c_i)$$

$$Entropy = -\sum_{i=1}^{n} p(c_i) log_2(p(c_i))$$

Where $p(c_i)$ is the probability of class i being in a node, and n is the number of classes in the target variable. Gini index and entropy are the criterion for calculating information gain. Decision tree algorithms use information gain to split a node. Entropy in statistics is analogous to entropy in thermodynamics and it signifies disorder. If there are multiple classes in a particular node, then that node is disordered. Information gain is the entropy of parent node minus sum of weighted entropies of child nodes. Weight of the entropies are number of samples in the node divided by total samples in the child nodes. Similarly, information gain can be calculated using Gini Index. The gini impurity measures the frequency at which any element of the data set will be mislabelled when it is randomly labelled.
The attribute selection measures are used to select the splitting attribute used at the node. The structure of a decision tree is shown below.

To demonstrate the effectiveness of the Decision tree classifier, we use the standard features of a car like the buying price, maintenance cost and luggage boot size to classify the safety level of cars as unacceptable, acceptable, good or very good. The pertinent data has been collected by Marko Bohanec and Blaz Zupan [1]. In the following sections we will dive a little deeper into decision tree classifiers, visualize the data collected, analyze it and then perform our model to classify new data points.

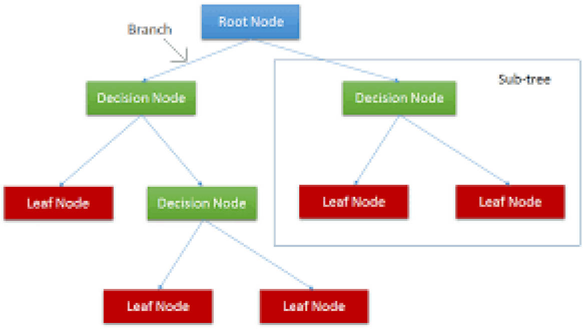


Fig. 16. The structure of a decision tree

Decision Trees

As mentioned in the section above, decision trees use the concepts of Gini Index and Entropy to classify between the multiple classes and designate mathematical rules for each class. There are two main differences between the Gini Index and entropy:

- The range of gini index is $[0, 0.5]$, whereas the range of entropy is $[0, 1]$. Maximum entropy and gini will be when both the classes in a binary classification have a probability of 0.5 for being in a particular node. Minimum gini and entropy are when one node contains only one class, i.e. the node is pure. For the best classification in a node, we would ideally want the minimum gini/ entropy depending on the parameter we use. The distribution of gini, entropy with probability of a class being in a node is shown below.

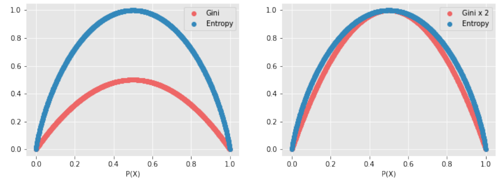


Fig. 17. The distribution of gini/ entropy with probability of a class

- Computationally, since the entropy requires algorithms, it is more expensive than calculating gini index. Calculating gini index is faster.

We will now look at some important advantages, disadvantages and metrics of evaluation for decision trees over other alternate machine learning algorithms that can be used for classification/ regression:

- Advantages
  1) Decision trees require much lesser data preparation than other supervised learning methods.
  2) Decision trees do not require normalization of data
  3) Decision trees do not require scaling of data too.
  4) Missing values in a decision tree do not affect the process of building a tree too.

5) Decision trees are easier to explain and visualize than most machine learning algorithms.

- Disadvantages:
  1) A small change in the data can change the entire structure of a tree. Very volatile, and high variance in expectations.
  2) Calculations can turn out to be much more complex than other algorithms.
  3) Takes higher time to train the model.
  4) Decision tree algorithm is inadequate for regression and in determining continuous values.
- Metrics of evaluation: As for many other machine learning algorithms, we can extract the classification report for our classification. The main metrics from the classification report that are of our interest are:
  1) Precision: How many values predicted to be in a certain class are in that class
  2) Recall: How many values in each class were given the correct label
  3) F1-score: weighted average of precision and recall

  The metric which you give the most importance to depends on your interest in false positives/ false negatives.

Pre-processing and Exploratory Analysis

In this section, we will first pre-process the data to extract the important features in the data provided, then we will visualize the data to understand the effects of certain features and modify the data-set to optimize the results from our decision trees model.

Pre-processing the data

The initial data-set has 1727 samples and 6 independent features, namely: the cost of buying the car, the cost for maintenance, number of doors in the car, number of persons that can sit in the car, how large is the luggage boot and predicted safety of the car. . Pre-processing the data-set includes (Note that the changes are made in both the training and testing data-sets):

1) Missing value assessment: No features in our data set have missing values, hence we can skip this part of pre-processing.
2) Additional Variables: No features seem to depend on each other and hence, it is safe to include every feature in our model.

Visualization

After preprocessing the data, we visualize the data, to better understand the correlation between different features:

1) Exploration of the countplots: The independent variables are almost unskewed, with an almost equal distribution of the counts of all the independent variables. The dependent variable seems to show a large number of ’unacc’ values, meaning that most of the cars are unacceptable with respect to their features.

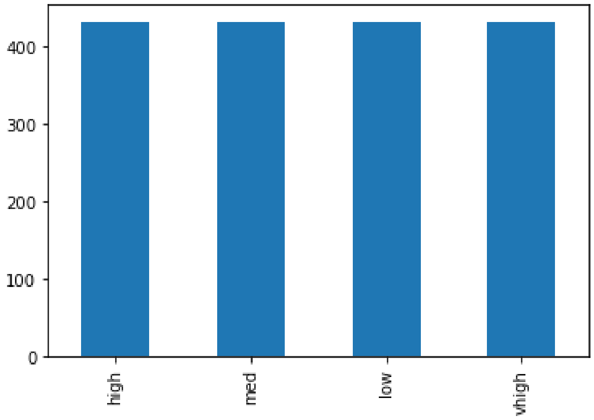


Fig. 18. Count plot of the ’buying’ column in the data set

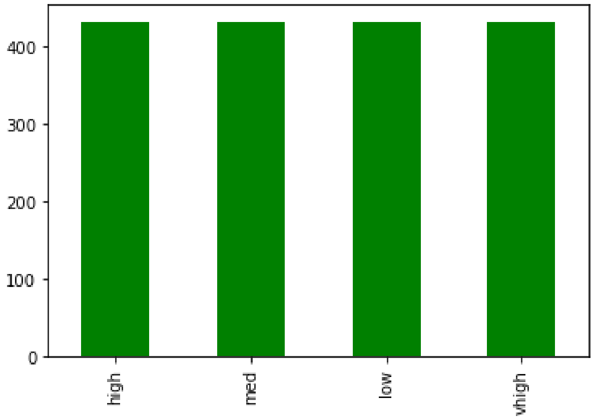


Fig. 19. Count plot of the ’maintenance’ column in the data set

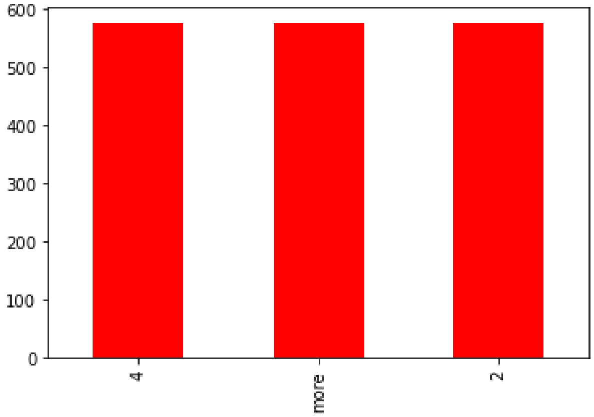


Fig. 20. Count plot of the ’persons’ column in the data set

2) Countplot of the luggage boot size per target: As the luggage boot size increases, the acceptance rate for the cars increases. The number of cars that are unacceptable are similar in number, but decrease as the luggage boot size increases.
3) Countplot of the luggage boot size per target: All cars

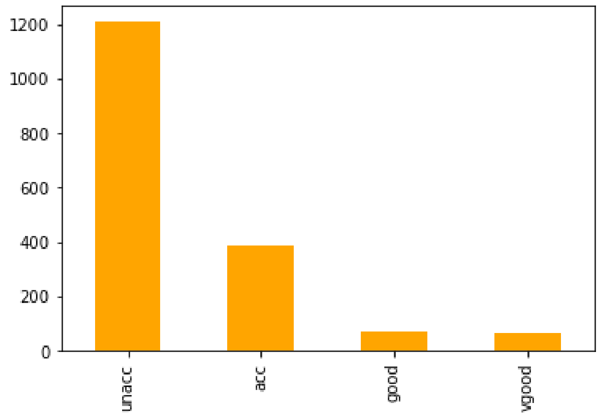


Fig. 21. Count plot of the 'target column in the data set

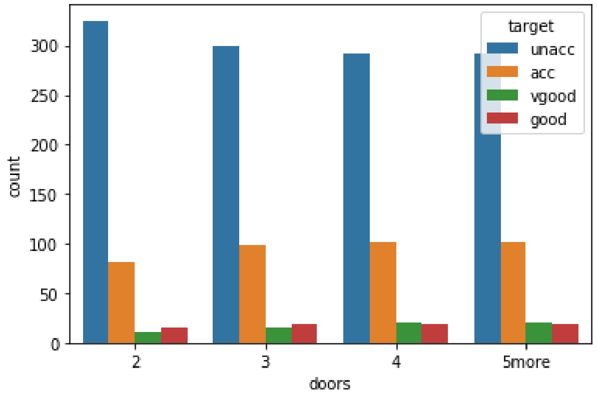


Fig. 22. Count plot of the number of doors per target

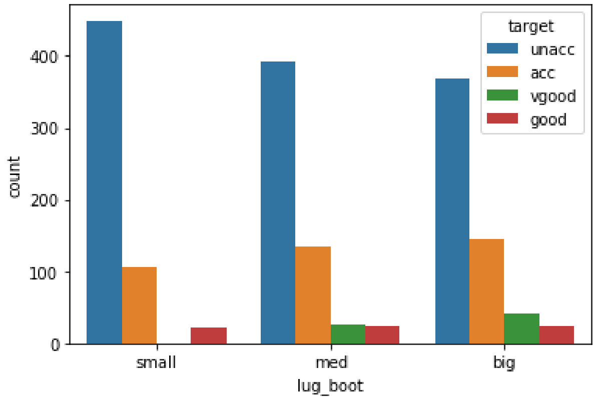


Fig. 23. Count plot of the luggage boot size per target

with a low predicted safety rating are unacceptable and should be rejected. More the predicted safety, higher the acceptance rate amongst those cars.

4) Trend amongst the cost for unacceptable cars: Surprisingly, most cars which are unacceptable cost really large amounts of money, and the acceptance rate increases as the cost decreases.

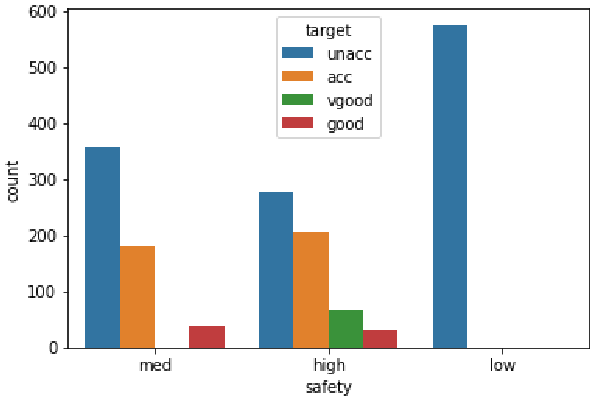


Fig. 24. Plot to compare the predicted safety levels and the target values

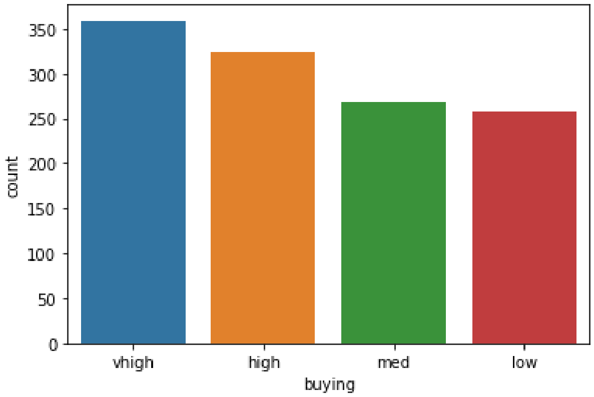
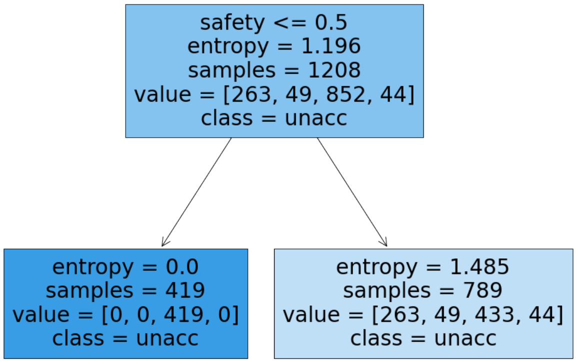


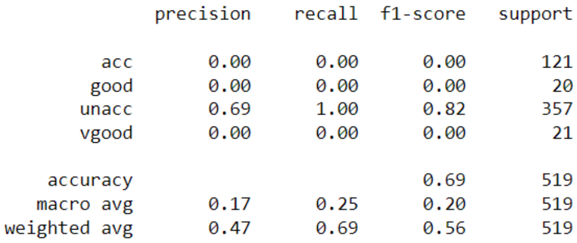
Fig. 25. Count plot of the buying column, when the target column is set to 'unacc'

The Decision Trees Model

After preprocessing the data, we move to the heart of the problem: the prediction of whether an unseen car (not in the training data-set) should be accepted. Here are the steps for the same:

- Appropriate Data type: Before applying a decision trees model to our data, we will analyze the data types for all variables, and change them, when necessary. Initially, all variables are of the categorical category. The scikit-learn package used to build the model cannot use categorical variables, and hence we will have to undergo a conversion process. Each variable follow a rank and order and thus, we perform an ordinal encoding on the data to maintain the order. For instance, the size of the luggage boot has three categories: small, medium, and large which has an order.
- Decision model and results: After the conversion of the data types, we run the decision trees model using the simple test/ train model, using the two criterion for classification: entropy and gini index and evaluate our results using the classification report. The decision tree we predict can also be visualized:

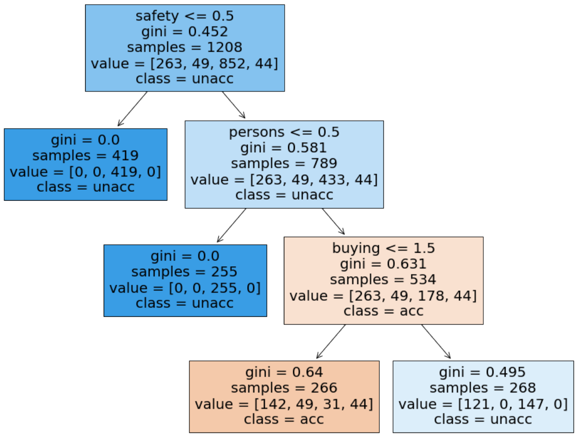


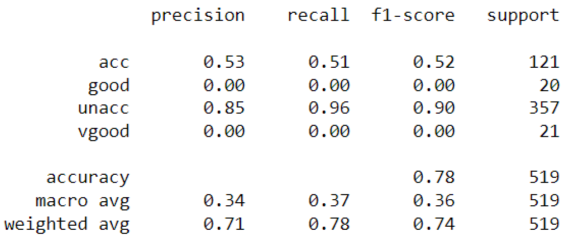


```
              precision    recall  f1-score   support

         acc       0.00      0.00      0.00       121
        good       0.00      0.00      0.00        20
       unacc       0.69      1.00      0.82       357
       vgood       0.00      0.00      0.00        21

    accuracy                           0.69       519
   macro avg       0.17      0.25      0.20       519
weighted avg       0.47      0.69      0.56       519
```

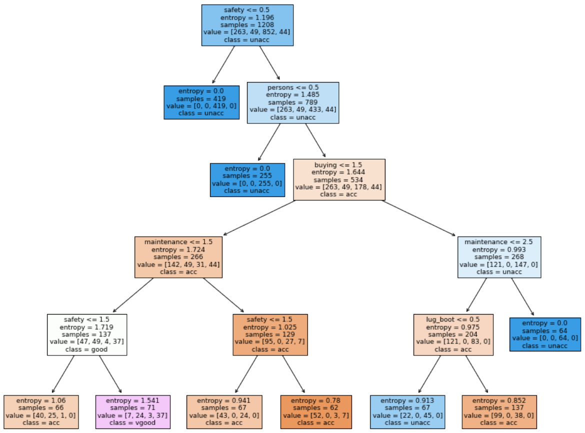


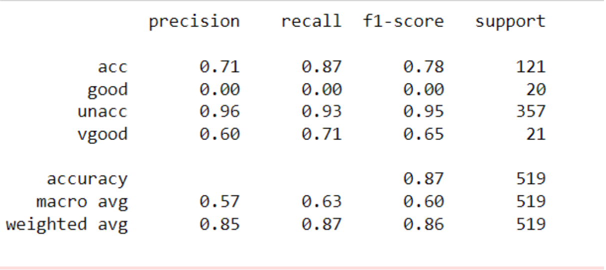


```
              precision    recall  f1-score   support

         acc       0.53      0.51      0.52       121
        good       0.00      0.00      0.00        20
       unacc       0.85      0.96      0.90       357
       vgood       0.00      0.00      0.00        21

    accuracy                           0.78       519
   macro avg       0.34      0.37      0.36       519
weighted avg       0.71      0.78      0.74       519
```

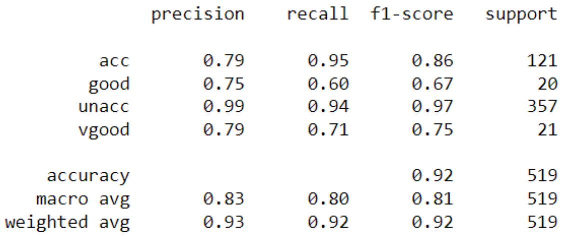




```
              precision    recall  f1-score   support

         acc       0.71      0.87      0.78       121
        good       0.00      0.00      0.00        20
       unacc       0.96      0.93      0.95       357
       vgood       0.60      0.71      0.65        21

    accuracy                           0.87       519
   macro avg       0.57      0.63      0.60       519
weighted avg       0.85      0.87      0.86       519
```



```
              precision    recall  f1-score   support

         acc       0.79      0.95      0.86       121
        good       0.75      0.60      0.67        20
       unacc       0.99      0.94      0.97       357
       vgood       0.79      0.71      0.75        21

    accuracy                           0.92       519
   macro avg       0.83      0.80      0.81       519
weighted avg       0.93      0.92      0.92       519
```

- The results are the same for the two classification criterion: gini index and entropy. However, as we change the maximum depths of the tree from 1,3,5,7 we see different results: the tree with max depth=7 gives the most accuracy and the tree with max depth 1 gives the least accuracy. Other metrics including precision, recall and f1-score all improve as the max depth of the tree increases.
- In this report, we can see that the model performs better for the unacceptable car samples than the acceptable cars, which is an interesting finding. Moreover, the good and very good sampled cars have zero precision, recall and f1-score. This means that the true positives, and the false negatives for the two conditioned cars are both zero. The two classes could have been avoided for better classification results.

Conclusion

The decision trees model was successfully demonstrated, both visually and numerically. The classification would have been better without the unneccesary classes, ’vgood’ and ’good’. As the max depth increases, the model becomes better.

V. Random Forests

***Abstract*—This paper aims to develop an understanding of the mathematical properties of random forests and showcase its application in classifying a car based on its safety. Number of trees in the random forest were changed and analyzed.**

Introduction

Random Forests is one of the most important tools used for classification and prediction. Once we receive data with some independent variables and a target variable which we want to determine for a given data point, we use statistical tools to develop a relationship between the target and the independent variables to decide on mathematical rules to assign a given data point to a class. The random forest classifier uses multiple decision trees for the classification.

Random Forests have applications spanning multiple fields. Like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction. The figure below depicts the same.

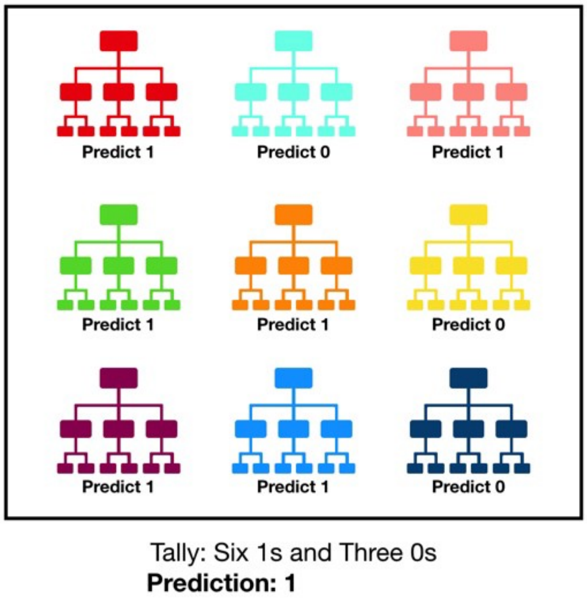


Fig. 33. Basic intuition of random forests

The fundamental concept behind random forest is a simple but powerful one — the wisdom of crowds.A large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models, and hence random forests are such a powerful method. The reason for this wonderful effect is that the trees protect each other from their individual errors (as long as they don't constantly all err in the same direction). While some trees may be wrong, many other trees will be right, so as a group the trees are able to move in the correct direction.

So how does random forest ensure that the behavior of each individual tree is not too correlated with the behavior of any of the other trees in the model? It uses the following two methods:

- **Bagging**: Decisions trees are very sensitive to the data they are trained on — small changes to the training set can result in significantly different tree structures. Random forest takes advantage of this by allowing each individual tree to randomly sample from the data set with replacement, resulting in different trees. This process is known as bagging.
- **Feature Randomness**: In a normal decision tree, when it is time to split a node, we consider every possible feature and pick the one that produces the most separation between the observations in the left node vs. those in the right node. In contrast, each tree in a random forest can pick only from a random subset of features. This forces even more variation amongst the trees in the model and ultimately results in lower correlation across trees and more diversification.

Random Forests

As mentioned in the section above, random forests use multiple trees made from random features and random data using bagging and feature randomness. In this section we will discuss some advantages, disadvantages, and the metrics of evaluation for random forests:

- Advantages
  1) Random Forests reduces the over fitting problem in decision trees and also reduces the variance and therefore improves the accuracy.
  2) Random Forest works well with both categorical and continuous variables.
  3) Random Forest can automatically handle missing values.
  4) No feature scaling is (standardization and normalization) required in case of Random Forest as it uses rule based approach instead of distance calculation.
  5) Random Forest is usually robust to outliers and can handle them automatically.
  6) Random Forest handles non-linear parameters efficiently.
- Disadvantages:
  1) Random Forest creates a lot of trees and combines their outputs. To do so, this algorithm requires much more computational power and resources. On the other hand decision tree is simple and does not require so much computational resources.
  2) Longer Training Period: Random Forest require much more time to train as compared to decision trees as it generates a lot of trees (instead of one tree in case of decision tree) and makes decision on the majority of votes.
- Metrics of evaluation: As for many other machine learning algorithms, we can extract the classification report for

our classification. The main metrics from the classification report that are of our interest are:

1) Precision: How many values predicted to be in a certain class are in that class
2) Recall: How many values in each class were given the correct label
3) F1-score: weighted average of precision and recall
4) Accuracy: This score measures how many labels the model got right out of the total number of predictions. Accuracy is not a great measure of classifier performance when the classes are imbalanced. We need more information to understand how well the model really performed.

The metric which you give the most importance to depends on your interest in false positives/ false negatives.

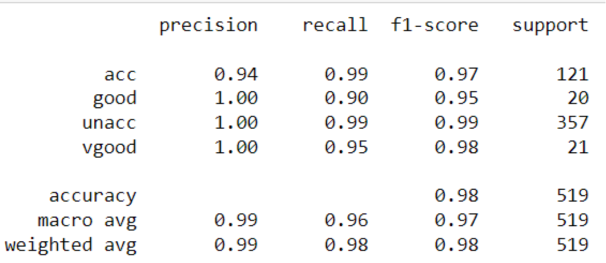
Pre-processing and Exploratory Analysis

The data used for both decision trees and random forests are same and hence the preprocessing done will also be the same. Hence, the details for the same are not mentioned here. Data visualization will also be the same for both the machine learning models.

The Random Forest Model

After preprocessing the data, we move to the heart of the problem: the prediction of whether an unseen car (not in the training data-set) should be accepted. Here are the steps for the same:

- Appropriate Data type: Data has been manipulated the same way it was done for decision trees.
- Random Forest model and results After the conversion of the data types, we run the random forests model using the simple test/ train model and evaluate our results using the classification report. We run the random forests model for three different number of trees: 100,200 and 500 trees. The classification reports for the different hyperparameters (number of trees) can be viewed below:



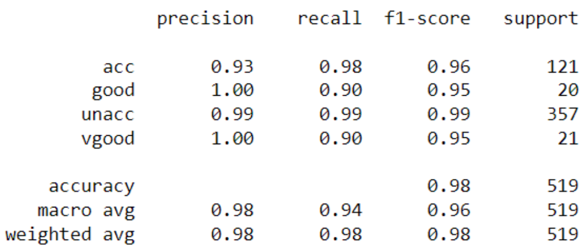
```
              precision    recall  f1-score   support

         acc       0.94      0.99      0.97       121
        good       1.00      0.90      0.95        20
       unacc       1.00      0.99      0.99       357
       vgood       1.00      0.95      0.98        21

    accuracy                           0.98       519
   macro avg       0.99      0.96      0.97       519
weighted avg       0.99      0.98      0.98       519
```

Fig. 34. Classification report with 100 trees

As shown by the results, random forests performs the classification excellently and delivers great results. The number of trees does not change the model a lot and each of the three decision trees give excellent results. The overall accuracy is 98% for the three models.

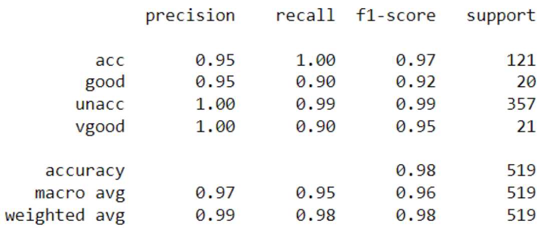


```
              precision    recall  f1-score   support

         acc       0.93      0.98      0.96       121
        good       1.00      0.90      0.95        20
       unacc       0.99      0.99      0.99       357
       vgood       1.00      0.90      0.95        21

    accuracy                           0.98       519
   macro avg       0.98      0.94      0.96       519
weighted avg       0.98      0.98      0.98       519
```

Fig. 35. Classification report with 200 trees



```
              precision    recall  f1-score   support

         acc       0.95      1.00      0.97       121
        good       0.95      0.90      0.92        20
       unacc       1.00      0.99      0.99       357
       vgood       1.00      0.90      0.95        21

    accuracy                           0.98       519
   macro avg       0.97      0.95      0.96       519
weighted avg       0.99      0.98      0.98       519
```

Fig. 36. Classification report with 500 trees

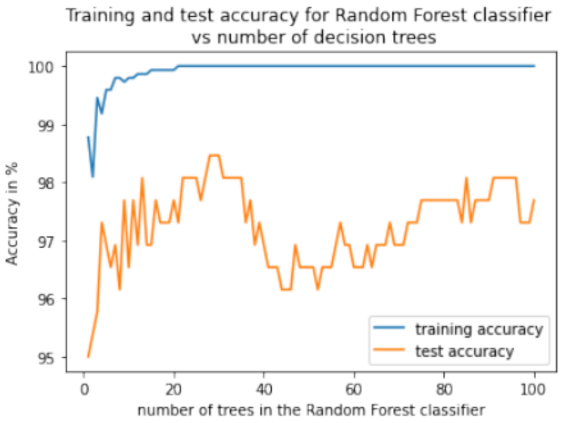


Fig. 37. Training and test accuracy of random forest classifiers against number of decision trees.

Conclusion

The random forest model was succesfully demonstrated, both visually and numerically. While the accuracy of the model is good enough, we could perform feature selection using feature scores.

References

[1] ”Sinking of the Titanic”, wikipedia.org, 2022 [Online]. Available: **Link**.
[2] ”Logistic Regression”, scikit-learn.org, 2022 [Online]. Available: **Link**.
[3] ”Car Evaluation Data Set”, archive.ics.uci.edu, 1990 [Online]. Available: **Link**.
[4] ”Sinking of the Titanic”, wikipedia.org, 2022 [Online]. Available: **Link**.
[5] ”Logistic Regression”, scikit-learn.org, 2022 [Online]. Available: **Link**.

VI. Exploratory Analysis of Stock Market Data

Introduction

The past few papers covered several aspects of Data Science. As a student of the field, it is very exciting to learn the applications of these techniques in real life. Having a personal interest in financial markets, applications of data sciences in this field was particularly interesting to me. Many decisions in the financial world are based on the data, hence data science play a crucial role in this field. Many of us are interested to invest in stocks or day to day trading, one of the most important steps for an investor or d2d trader is to visualize a stock's performance and take a decision.

We begin with the visualization of the stock's data and attempt to judge the stock's performance in the market. Visualizations act like a human intuition for investing in stocks.There are other sophisticated financial metrics like risk to benefit ration which act like a stronger proof for someone to invest in a stock. We attempt to predict the stocks closing price in the near future with simple lasso regression.This prediction will make the investor more confident in their moves.

This paper aims to perform an exploratory analysis of stock market data and also to help improve an investor's behaviour. We use techniques of Data Science to analyse the market and influence investor's decisions.

Visualizations

We will begin our quest to determine stock prices by visualizing the stock data provided to us. The information that has been provided to us are: day-wise stock closing, opening, highest and lowest prices. The data included stocks of six companies.

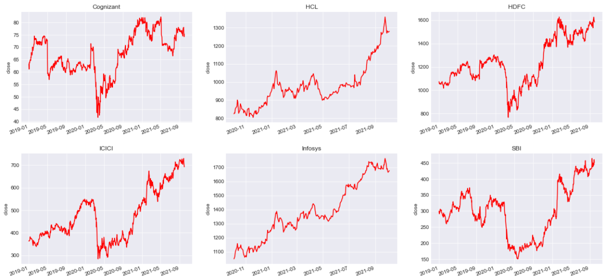


Fig. 38. Close prices of the six stocks

1) Exploration of the closing prices: To understand the stocks performance over a period of time we can look into the plots of closing prices. By looking at these plots we can come to a preliminary conclusion of whether the stock is bullish or bearish.
2) Exploration of the candlestick plots: Candlestick plots are popular in stock market analytics. A candlestick includes the closing, opening, high and low price of the stock during a particular time period. candlestick plots are very important graphics for an investor's decision making process. The patterns and their implications are beyond the scope of this paper.

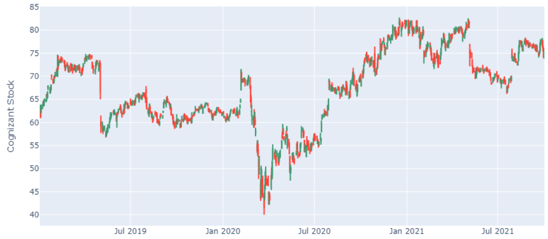


Fig. 39. Candlestick plot for Cognizant stock prices

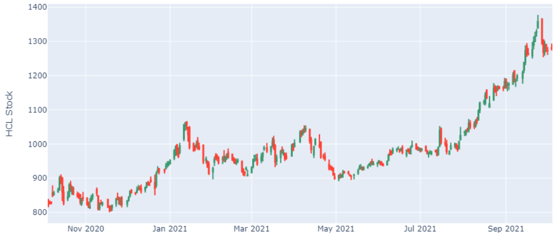


Fig. 40. Candlestick plot for HCL stock prices

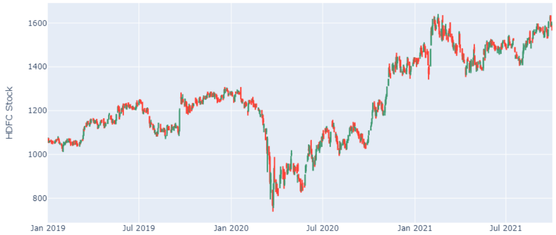


Fig. 41. Candlestick plot for HDFC stock prices

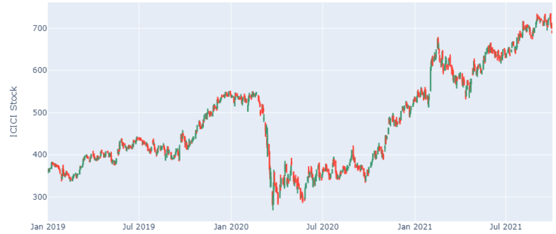


Fig. 42. Candlestick plot for ICICI stock prices

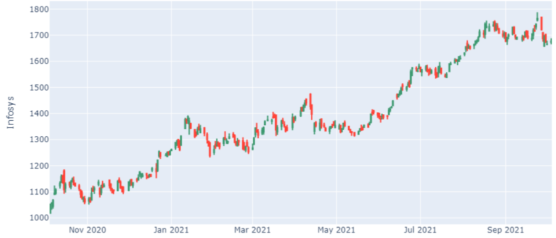


Fig. 43. Candlestick plot for Infosys stock prices

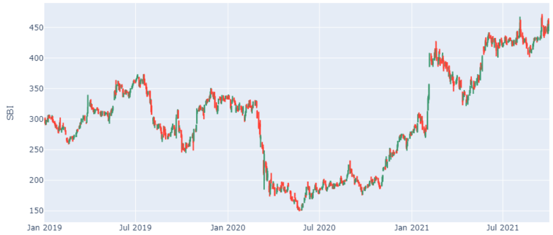


Fig. 44. Candlestick plot for SBI stock prices

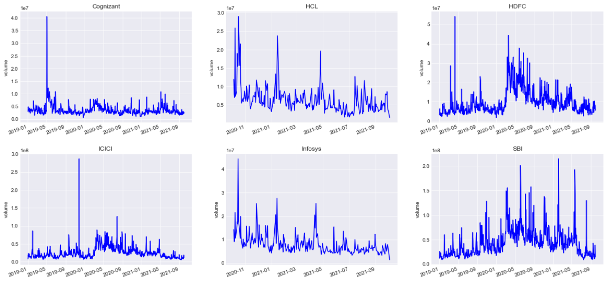


Fig. 45. Plot for the volumes of stock bought

3) Exploration of the volume of stocks bought: Another important aspect of stocks is the volume in which it is being trading. It directly reciprocates to the popularity of the stock.The spikes in the plot could imply a major financial event for the company. Volume is a great metric for investors to explore before making a decision when the rest of the metrics are not reliable.

Predicting the closing stock prices

After preprocessing the data, we move to the heart of the problem: the prediction of closing stock prices for the six given companies, using three different linear regression models as described below.

TABLE I
Models' Results

Company	OLS	Ridge	Lasso
Cognizant	0.8701	0.8701	0.7875
HCL	0.9294	0.9294	0.9295
HDFC	0.9175	0.9175	0.9170
ICICI	0.9314	0.9314	0.9355
Infosys	0.8937	0.8937	0.8858
SBI	0.9730	0.9730	0.9680
Average	0.9192	0.9192	0.9038

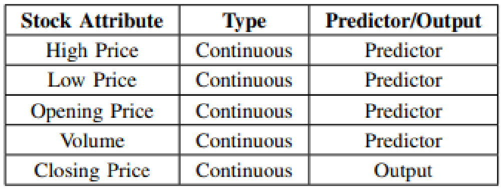
A. Regression Models

The three different regression models (OLS Linear Regression, Ridge Regression, Lasso Regression) are explained below:

- OLS Regression: Ordinary least squares (OLS) is a type of linear least squares method for choosing the unknown parameters in a linear regression model by the principle of least squares: minimizing the sum of the squares of the differences between the observed dependent variable in the input dataset and the output of the (linear) function of the independent variable.
- Ridge Regression: Ridge regression is just a improvement on ordinary least squares regression. In ordinary least squares regression the objective is to minimize the residual sum of squares (RSS). In ridge regression we add a penalty to the objective function. The penalty is sum of square of values of coefficients.
- Lasso Regression: Lasso regression is very similar to ridge regression. The penalty is sum of absolute of values of coefficients.

B. Data Provided

Before comparing the three models on the stock data, we will have a look at the data given to us. We have 5 columns of features available to us for the six different companies, and each of these features are continuous. We can use the closing price as the output and each of the other features as predictors since predicting the closing prices over a given time is of most use in the real world.



Stock Attribute	Type	Predictor/Output
High Price	Continuous	Predictor
Low Price	Continuous	Predictor
Opening Price	Continuous	Predictor
Volume	Continuous	Predictor
Closing Price	Continuous	Output

Fig. 46. Summary of the data provided

C. R2 scores for each regression model

Table 1 shows our results of the R2 scores for the three different models on the six stock prices.

D. Visualizing the predictions

From the results of the R2 table, we realise that all the regression models chosen perform similarly and therefore we use the simplest model: OLS to predict the future stock prices of the companies given to us.

To perform our predictions, we have split our day-wise data up to a particular date, such that the train data include 95of the data and the rest of the data or the validation data set is used for prediction of future prices. Regression models do not perform well on time series data, while few deep learning models are specifically made for such type of data. We can look at the R2 of the models to understand the performance of the OLS regression models.

We can see that the regression model performs decently well on three of the stocks, while the low R2 value in one stock could mean that data was not sufficient or that the model's behaviour could not be mimicked by a linear model.

The visualizations for the same are provided below.

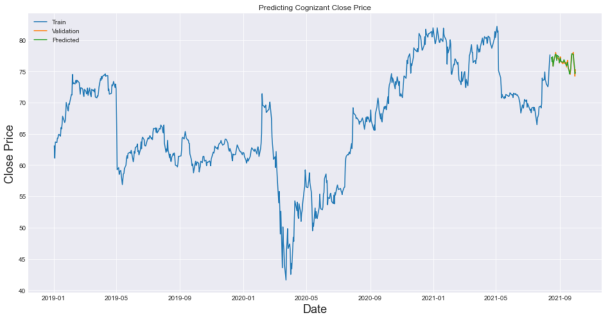


Fig. 47. Stock Price prediction for Cognizant

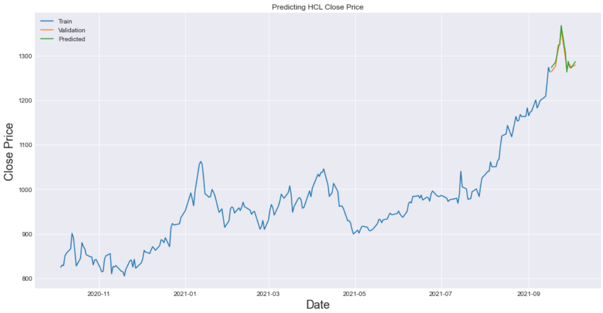


Fig. 48. Stock Price prediction for HCL

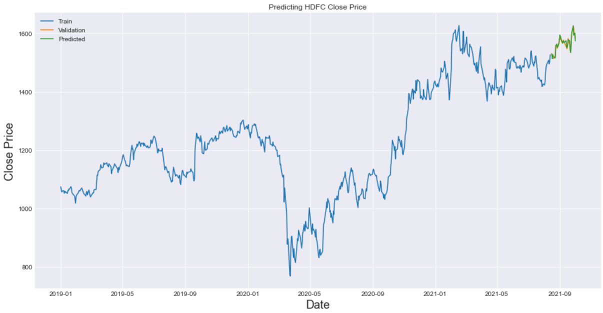


Fig. 49. Stock Price prediction for HDFC

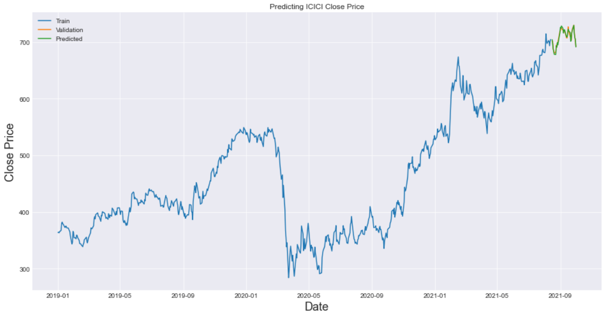


Fig. 50. Stock Price prediction for ICICI

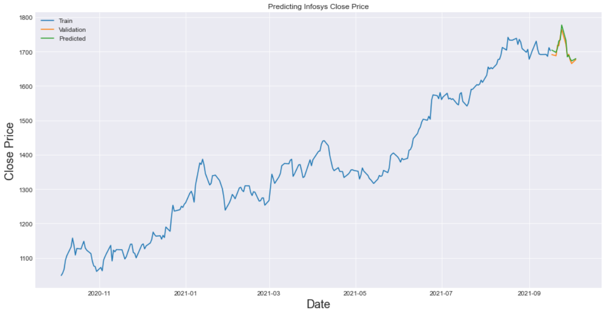


Fig. 51. Stock Price prediction for Infosys

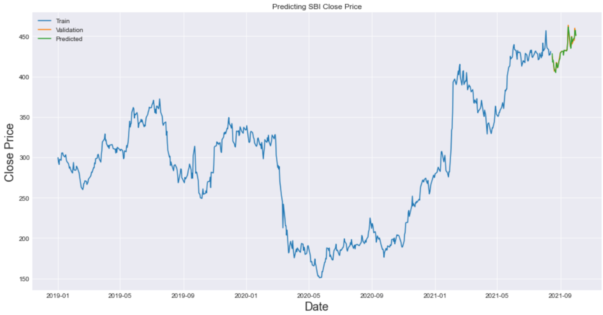


Fig. 52. Stock Price prediction for SBI

Conclusions

The impact of Data Science on Financial world is significant, every monetary decision involves the usage of analytics as a support system. This paper depicted one such usage of statistical learning models in financial world. Regression is often overlooked for complex problems like the stock market analysis this paper focuses on. While these systems may not follow a linear behaviour, regression could help us understand the trends of the data. Sometimes we do not look for accuracy but for interpretability.

References

[1] ”Japanese Candlestick Guide”, Patrick Foot, 2022 [Online]. Available: **Link**.

[2] ”An Introduction to Statistical Learning : with Applications in R”, Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, 2022 [Online].