A comparative study of the prediction of logistic regression, Regression Trees and Random Forest.

I. Introduction:

The motivation behind this project is to compare the performance of logistic regression with that of the regression tree method and Random foresting and to try and estimate which would provide better results given certain data. The base reference for this type of data is the paper that I have chosen is - '***Mueller, C. W., & Price, J. J. (1990). Economic, psychological, and sociological determinants of voluntary turnover. Journal of Behavioral Economics, 19(3), 321–335***'. My aim is to study the prediction performance for each model based on the calculated Mean Squared Errors of Logistic Regression, Regression Trees and Random Forest based out of the data (but not a one to one replication of the data from the paper)from '***Mueller, C. W., & Price, J. J. (1990). Economic, psychological, and sociological determinants of voluntary turnover. Journal of Behavioral Economics, 19(3), 321–335***' and try to provide an analysis of the models performances.

II. A brief look into the paper:

The paper aims to study voluntary turnover amongst employees. The factor that makes this paper unique and stand out from other existing literature during the time of publishing is that apart from the standard economic factors of measuring turnover via the cost-benefit-analysis, this paper also implements psychological and sociological factors(which can be related behavioral-related aspects) into the possible causal estimation of voluntary employee turnover. For this purpose, the authors take data from a cohort of 135 then recently hired registered nurses who were employed by a university hospital and were analyzed to assess the effects of the various explanatory variables on turnover during one year of employment. Then, turnover is measured by organization records for 12 months following the administration of the questionnaire designed to measure the independent variables.

III. Data:

The data was collected from a study of all registered nurses who began working for a large mid-western hospital in the United States Of America between the months of June 1983 and October 1984. Each month, the new entering cohort of nurses were brought into the study. In a total, three hundred and fifty nurses were asked to take part in the surveys. Surveys were administered again at two intervals, each six months apart from each other(at the time of entry of new nurses). These surveys were carried out to capture the explanatory variables, while the turnovers, voluntary or involuntary were recorded by the hospital. The authors attempted a possible explanation of the turnover of nurses during their sixth to eighteenth-month period(after their entry). The turnover behaviour is hence recorded over a period of twelve months.

IV. Methodology used in the paper:

The dependent variable in the study is "Turnover" . This dependent variable is dichotomous in nature. This is measured by hospital records and has the value of one for voluntary leaver and zero for an employee who stays and is measured in the twelve month period as mentioned earlier. For this purpose, the authors perform a logistic regression.

IV.A. Logistic Regression:

When dealing with data that has binary outcomes, that is variables that can have only two possible values: for instance - Yes or no, 1 or 0 (could also be 1 if yes, 0 if no) and so forth. For instance, in this case, when studying the cause of employee Turnover depending on several variables and where Turnover is in the form of 1 and 0. In such a case, typical regression models could predict off-scale values below zero or above 1 and hence it makes sense to model the probabilities on a transformed scale; this is what is done in logistic regression analysis. A linear model for transformed probabilities can set up as:

$$logitp = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

Where $logitp = [p/(1-p)]$ is the **log odds**.

In a logistic model, it is important to note that there is no error term as in linear models. This is because we are modelling the probability of an

event directly and this determines the variability of the binary outcome. There is also no variance parameter as in normal distributions.

A little side note which is relevant to the coding part that I have done is that logistic regression belongs to the class of '*generalized linear models'* . Such models are characterized by their response distribution(in this case, binomial distribution) and a certain link function, which transfers the mean value onto a scale where the relation to the background variables in the models is described(as linear and additive) . In a logistic regression, the above mentioned link function is the **log odds.**

IV.B. Methodology used as a part of my study:

1) Regression Tree:

A standard linear regression is a 'global model', where in we have a single predictive formula, which is used upon a certain data. But, when we are faced with situations where a certain data set has many features which interact in complex and non-linear ways, assigning a single 'global model' could be very taxing. Hence, an alternative to a non-linear way is to use a model which would sub-divide or classify(or partition) the space into smaller, more comprehendable regions. We then further partition these sub-divisions until we are left with small chunks of portions which are easy to analyze and where-in a model could be fit to them. This process is called '**recursive partitioning'.** Decision Trees use the tree to represent recursive partitioning. The tree has terminal nodes and leaves, each of which represent a cell of the partition and has to it attached, a simple model which applies exclusively to that cell. For instance, a point **x** belongs to a particular leaf, if it falls in the corresponding cell of the partition. To find out which cell we are currently in, we must start at the root node of the tree . The nodes that are on the interior are labeled with questions and the edges or branches between them carry labels with the answers.

The model for a regression tree is of the form:

$$y_i | x_i = \mu + \epsilon_i$$

2) Random Forest:

Random Forest is a method where the algorithm builds a 'forest', which is an ensemble of decision trees. One of the core features of this method is that it is trained on a data set(training) with the 'bagging' method. In other words, random forest constructs multiple decision trees and at the end, merges them to get an accurate and stable prediction. Random forest can be used for both classification and regression problems.
Random Forests overcome the problem of variance faced by the bagging method by forcing each split to consider only a subset of the predictor variables. This implies that on average, (p-m)/p of the splits will not even consider a stronger predictor(if present in the data set) and the other predictors will have more of a chance than in other methods. This process is often referred to as **'decorrelating'** the trees and therefore making the average of the trees with lesser variance and hence more reliable.

V. STUDY:

As a part of my study, I carry out a comparative analysis of the predictions of three different classification based predictors, Logistic Regression, Regression Tree and Random Forest. To start off with, I will briefly go through my data generating process. The data generating process that I have used is based out off the paper '***Mueller, C. W., & Price, J. J. (1990). Economic, psychological, and sociological determinants of voluntary turnover. Journal of Behavioral Economics, 19(3), 321–335'*** where in, the dependent variable - '*Turnover'* is of a binary nature, meaning the value is 1 if an employee voluntarily leaves and 0 if he/she stays. I set a threshold at 0.65 such that an observation above that is set to 1 and an observation below it is set to 0. The total number of observations is 1500, with a total of 21 independent variables, on which the dependent variable 'Turnover' is regressed on. I then split the data(1500 observations) into training and testing (75% = training and 25% = testing) to further carry out my study. The aim of my study is to see how different classifiers would perform on data generated in such a way. The key measure for comparison of performance that I have implemented is MSE(Mean squared error). For each of the three models, I train them on a training data set and then run the trained model on a test data set and then compare the results(MSE) of each models performance.

VI. Results and Conclusion:

After the data generating process is done, I move on to the models.

- **The first model is the Logistic Regression Model**. After training the model on the training data and then running it on the test data and then calculating the Mean Squared Error, the random generator gives a MSE of (circa) 1.154 which is a very small error rate and goes to show that the model performs well.

Further, We can also see the graph of the **confusion matrix.**

Confusion matrix: A **confusion matrix** is a table that can be used to interpret the performance of a classification model (or "classifier") on a set of test data.
From this matrix, we can see that the logistic regression model performs well gives good predictions for 'True Negatives' and 'True Positives' . Although there are some amount of 'False positives' and "False negatives' which amount to errors in the prediction, these are of an acceptable number.

- **The second model is the Regression Tree model:**

After following the same process of training the model on the training and running the model on the test data, I generate two different trees to get a better understanding on the prediction performance of the Tree model. The **first** is the full (unpruned) tree. The full tree has a total of 50 splits and the random generator shows that the calculated **MSE is (circa) 4.0928.**
The complexity parameter plot shows that the error rate slowly increases with a larger number of splits. We can see that around the 16^{th}-21st split, there is an observable climb in the error rate.
Second is the pruned tree. This tree has a total of 20 splits. The model decides to stop further splitting after this, as it decides that this is the optimum point, where the error is at the optimum level(as seen before, beyond this level, the error rate slowly starts to rise). The random generator calculates the **MSE to be (circa) 4.045.**

- **The third and final is the Random Forest method:**

I go on to do the similar process of training the model on a training set and then testing it on the test set and then calculate the MSE generated by

the model. The random generator shows a calculated **MSE of (Circa)0.176 .**
Then, I go on to also generating a confusion matrix which shows an even better result than that of the confusion matrix of the logistic regression. We see that the 'True positives' and the 'True negatives' have been predicted well, despite some negligible amount of errors(False positives and False negatives) .

For a **final** comparison, I run a simulation study wherein each model is trained over a training data set, then tested on a test data set and then the MSEs are calculated in a loop for a total of 100 times and with a slightly larger data set (previous data had a total number of observations of 1500 and now, 2000) to see how the models perform over time and with a larger data set. The results follow the same pattern as for when they were simulated only once(as done before) and Random Forest model has the lowest MSE, followed by Logistic Regression model and Regression Tree model. **The three MSEs being (circa) : 0.032, 0.239,0.761**, respectively.

From this study that I have done, I can conclude that with data that is suitable for a classifier such as logistic regression, Random forest still performs the best, followed by logistic regression and then regression trees. Although we can see that after running the final simulation study, over 100 times the average of the MSEs for all three methods seems very less(possible due to the law of large numbers) and goes to say that they are good predictors, we can see a slight difference in all three, which can be seen in a more magnified form when they are simulated for only one time.