A Project Report

on

OLS and Random Forest Regression using R

of

Business Analytics (BM516H)

Submitted by:

Debashish Rajbongshi (224024004)

Om Prakash Ahirwar (224024011)

Pritam Ghosh (224024013)

Under the guidance of:

Dr. Kuldeep Baishya and Dr. Abhay Pant

School of Business

Indian Institute of Technology, Guwahati

Objective:

To study the given financial and human resource data of employees of an organization to determine the new CTC for potential/new employees using a predictive regression model.

It has been observed that nowadays companies find it very difficult to predict the salaries of new or potential employees. So we have tried to make a predictive model so as to outline the most important variables to be considered while predicting the salary.

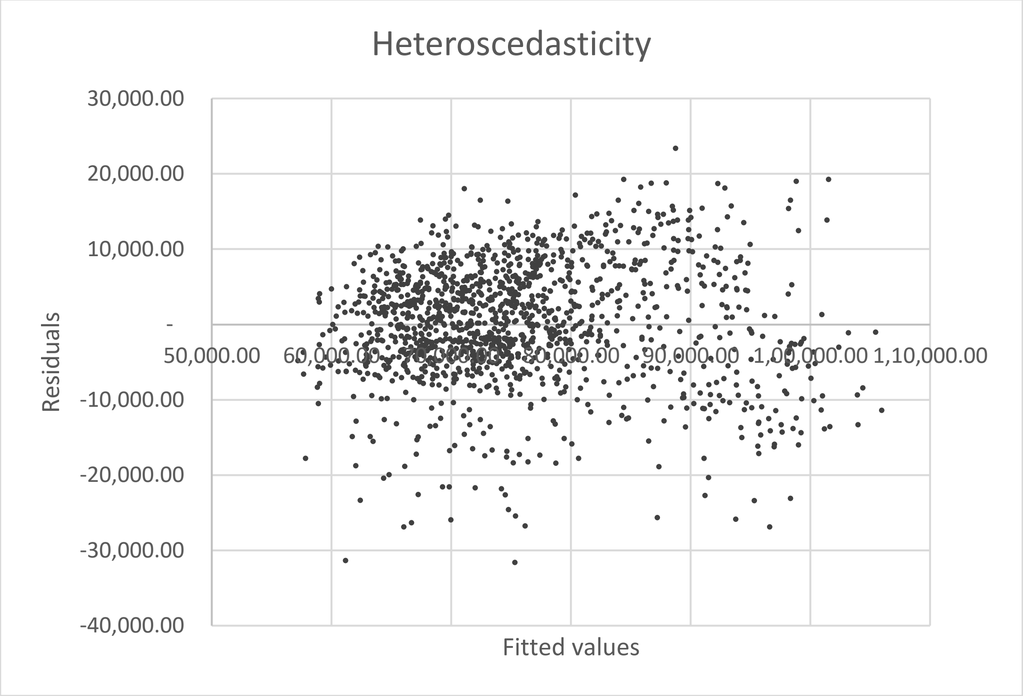
We have taken the HR data from a project in analytics provided by Internshala Trainings.

Here our dependent variable is predicted monthly CTC. There are seven independent variables involved.

These are College type, City type, Job role, Previous CTC, Number of Previous Job Changes, Graduation Marks, Experience (in months), out of which there are 3 categorical variables i.e., College Type, Job Role and City Type. These categorical variables have to be encoded i.e., they have to be converted into dummy variables. We performed one hot encoding for our categorical variables using R.

Our Approach:

We had previously done OLS regression on this data set using Excel. However, we experienced heteroscedasticity in this model, as explained by the graph as well as Breusch-Pagan test which rejected the null hypothesis of homoscedasticity. The calculated chi square value was much higher than the value from the table (90.48 > 15.5).



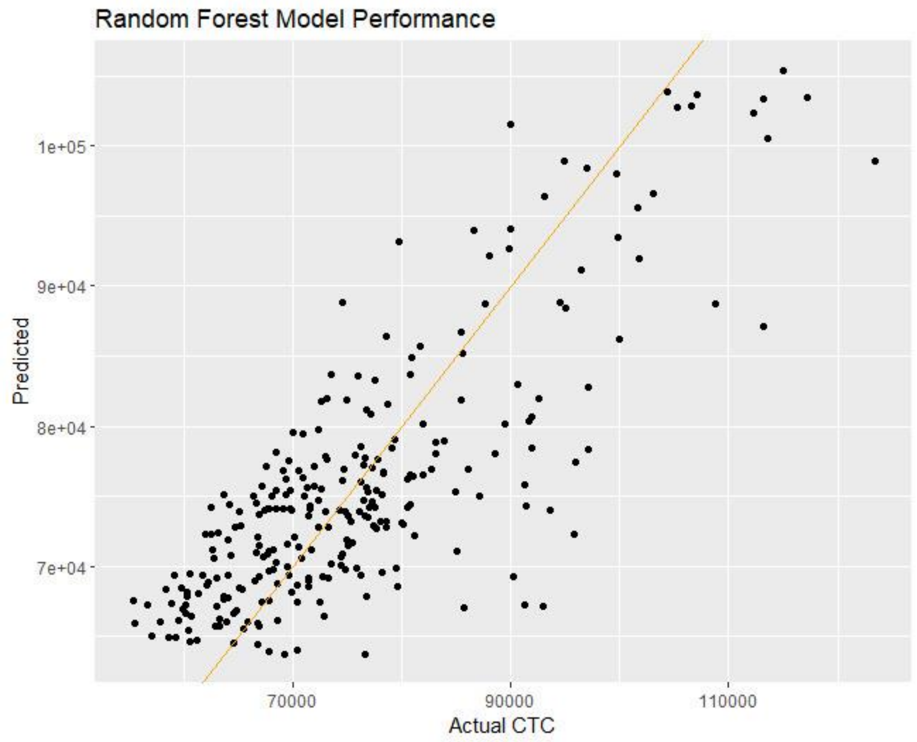
Heteroscedasticity in OLS regression model

There was no autocorrelation present as DW value came out to be 2.05.

Correlation matrix between independent variables showed relatively high correlation between the dummy variables College_Tier_1 and College_Tier_2. However, we could not conclude on multicollinearity being present as the VIFs we got were 1.55 and 1.63.

We also built a random forest model using R studio. Random Forest is generally considered robust to heteroscedasticity and multicollinearity because it is an ensemble learning method that combines multiple decision trees, and the effects of individual decision trees' errors are mitigated through the aggregation process. Additionally, the Random Forest algorithm can handle a wide range of data types and distributions, including non-normal distributions.

Model performance looks like this using line of perfect prediction:



Here is a scatter plot of Predicted CTC. Ideally, we would like the points to be scattered around the red line (line of perfect prediction), where Predicted CTC = Actual CTC in the train and test set.

R code for Random Forest:

```
library(readxl)

library(dummy)

library(caret)

library(randomForest)

library(ggplot2)

library(Metrics)

library(lmtest)

library(car)

HR_data <- read_excel("C:/Users/Administrator/Desktop/HR_data.xlsx")

View(HR_data)
```

```
cat_data<- data.frame(College=HR_data$College,

            Role=HR_data$Role,

            `City type`=HR_data$`City type`)

cat_data<- dummy(cat_data)

View(cat_data)

HR_data <- cbind(cat_data, HR_data[, -c(1:4)])

set.seed(123)

#Partitioning to training and test set

index <- createDataPartition(HR_data$CTC, p = 0.8, list = FALSE)

train <- HR_data[index, ]

test <- HR_data[-index, ]

View(train)

#For Random Forest

model <- randomForest(train$CTC ~ . , data = train, importance = TRUE, ntree = 1000)

test$Predicted_CTC <- predict(model, newdata = test)

View(test)

#To Calculate root mean square error

# Calculate MAPE

MAPE <- mape(test$CTC, test$Predicted_CTC)

# Calculate RMSE

RMSE <- rmse(test$CTC, test$Predicted_CTC)

MAE <-mean(abs(test$CTC-test$Predicted_CTC))

r_squared <- cor(test$Predicted_CTC, test$CTC)^2

#plot
```

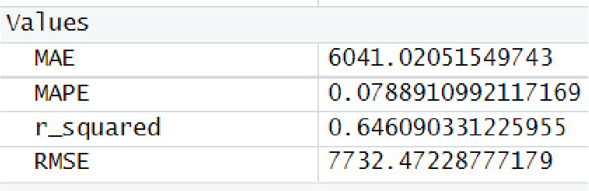
```
ggplot(test, aes(x = CTC, y = Predicted_CTC)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1, color = "orange") +
  labs(x = "Actual CTC", y = "Predicted", title = "Random Forest Model Performance")
#residual analysis
test$resid_rf<-test$CTC- test$Predicted_CTC
#Trial for new employee
new_test<-read_xlsx("C:/Users/Administrator/Desktop/test2.xlsx")
View(new_test)

new_test$Predicted_CTC <- predict(model, newdata = new_test)

importance(model)
varImpPlot(model)
```

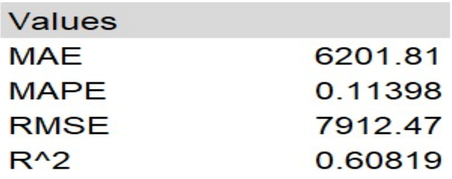
Comparison between our models:

Since random forest models do not provide any regression summary, we can still compare the random forest model with OLS regression model using measures such as MAE (mean absolute error), RMSE (Root Mean Square error) and MAPE (Mean Absolute Percentage error).



Values	
MAE	6041.02051549743
MAPE	0.0788910992117169
r_squared	0.646090331225955
RMSE	7732.47228777179

Random Forest



Values	
MAE	6201.81
MAPE	0.11398
RMSE	7912.47
R^2	0.60819

OLS Regression

We can see all the measures of Random Forest fare better than that of OLS Regression i.e., errors are significantly lower and R^2 is higher in random forest model. So we should adopt this model.

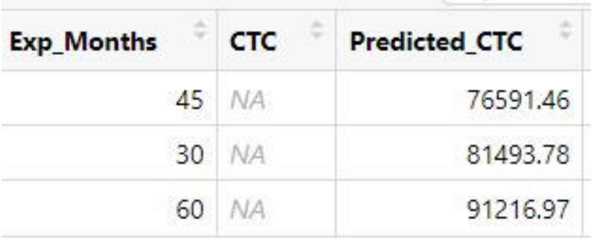
Prediction differences between the two models:

We have taken 3 potential employees whose salaries have to be predicted. Their respective profiles are as follows:

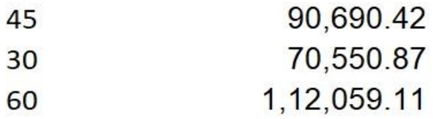
1. Tier 1 college, Manager role, Non-metro city, previous salary was 50000, 1 job change, graduation marks in 60% and 45 months work experience.

2. Tier 2 college, Executive role, Metro city, previous salary was 60000, 2 job changes, graduation marks in 65% and 30 months work experience.

3. Tier 3 college, Manager role, Metro city, previous salary was 90000, 4 job changes, graduation marks in 50% and 60 months work experience.



Exp_Months	CTC	Predicted_CTC
45	NA	76591.46
30	NA	81493.78
60	NA	91216.97



45	90,690.42
30	70,550.87
60	1,12,059.11

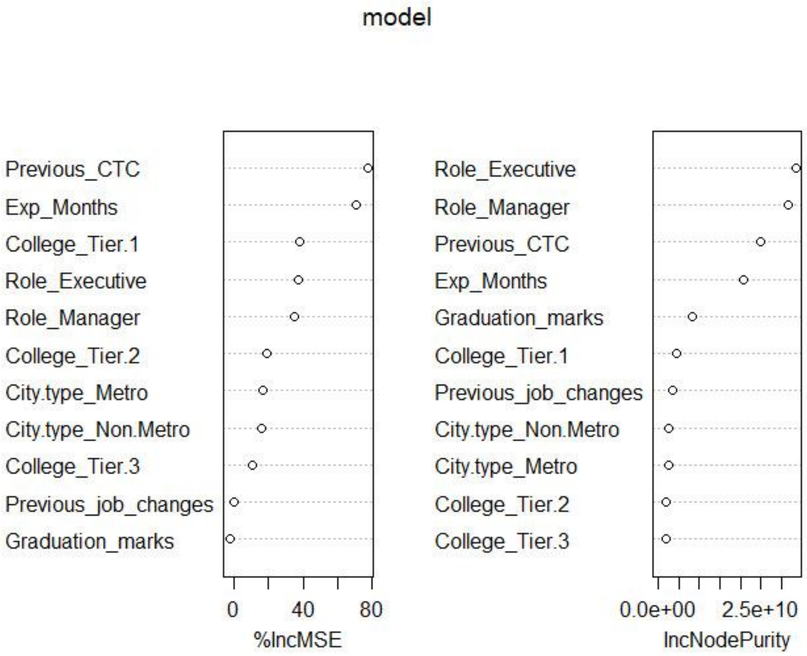
Random Forest

OLS Regression

In most cases, OLS regression model predicts higher salary figures than random forest model, hence one inference could be that this random forest can help the HR department to save on the bugdet allocated for manpower.

Inference:

As much as we can stress on the positive performance of random forest over OLS regression, we can also figure out which variables majorly influence the CTC of an employee.



Previous CTC :	77%
Experience in months :	70%
Role of the Job :	41%
College Tier :	25.6%
City Type:	19%
Previous job changes:	Close to Zero
Graduation Marks:	Close to Zero

Since the predicted salary is mostly influenced by factors like Previous CTC and Experience in months, these factors should be given the most priority by the HR department while determining the compensation and benefits offered to the potential employees.