Employee Attrition Prediction Analysis

Objective: To predict if an employee is going to resign or not

Methodology: 1. Through our analysis we intend to build a model which can predict if an employee is about to quit. 2. We shall be looking at all variables through some plots and infer about it in our exploratory analysis. 3. After our exploration we shall build some features based on the Variables at hand and take a call on inclusion/exclusion of few variables.

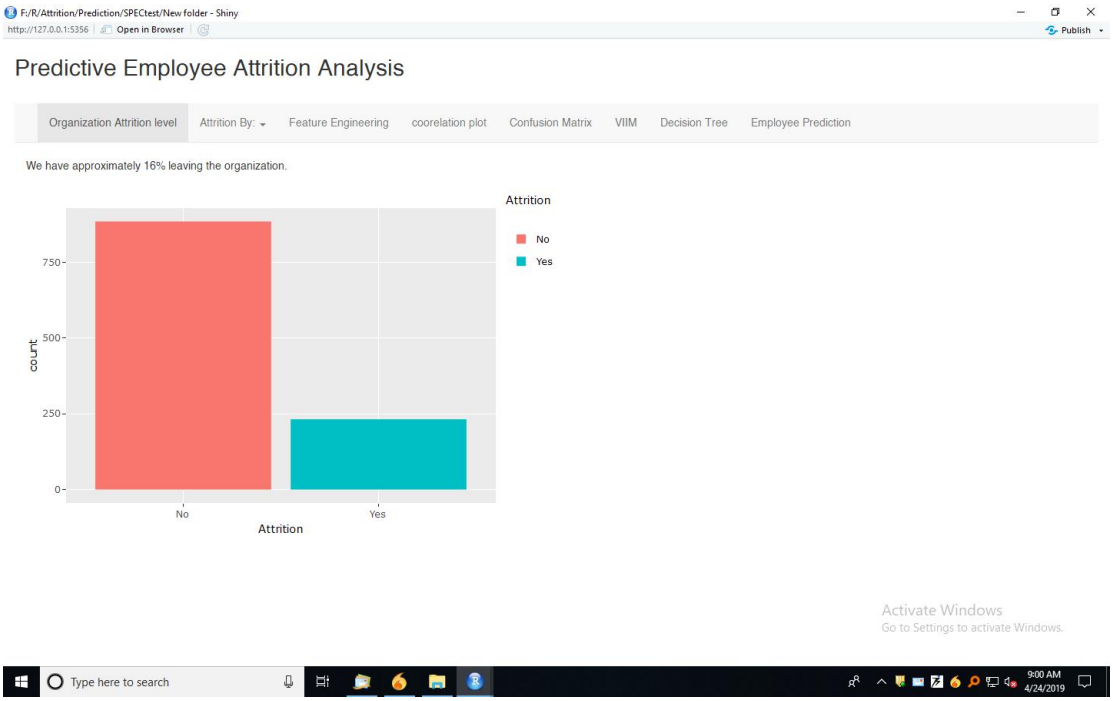


Fig Overall Attrition

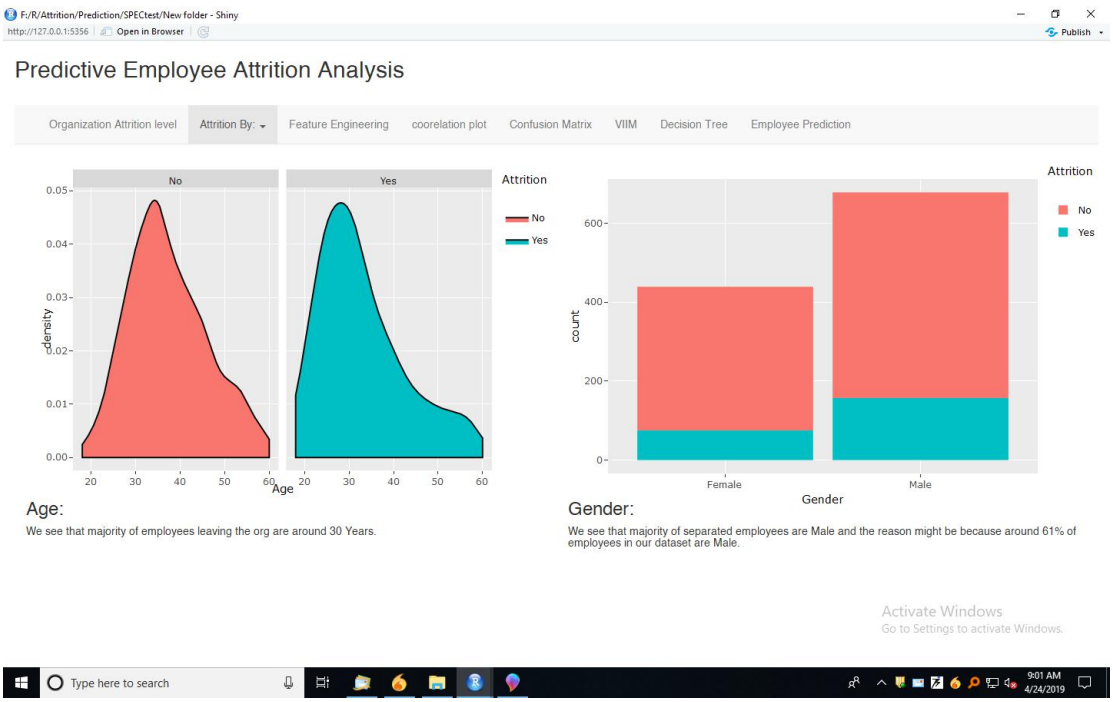


Fig Attrition by Age and Gender

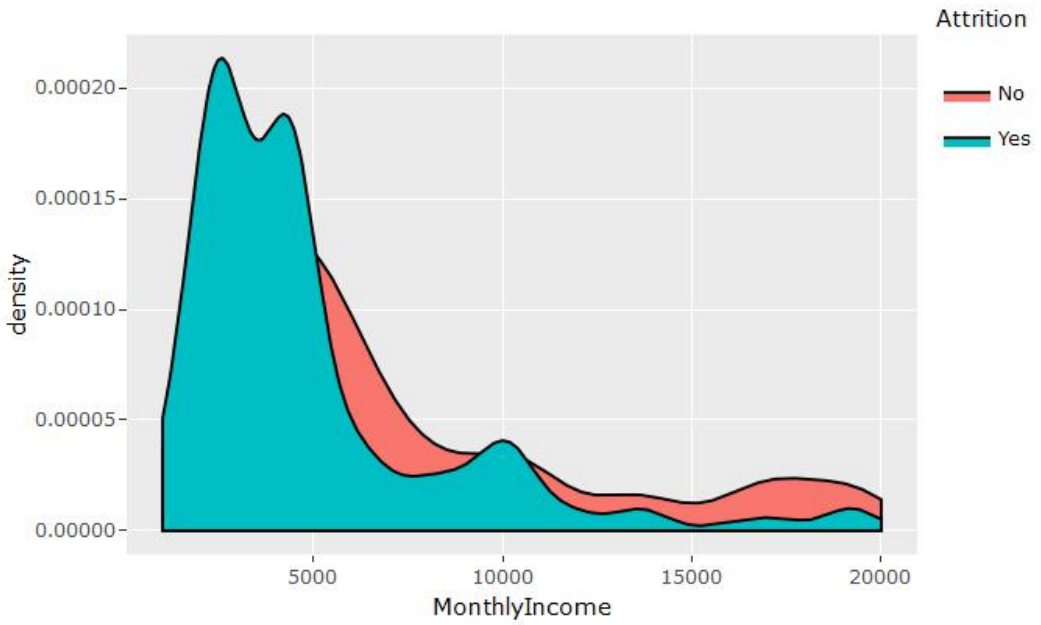
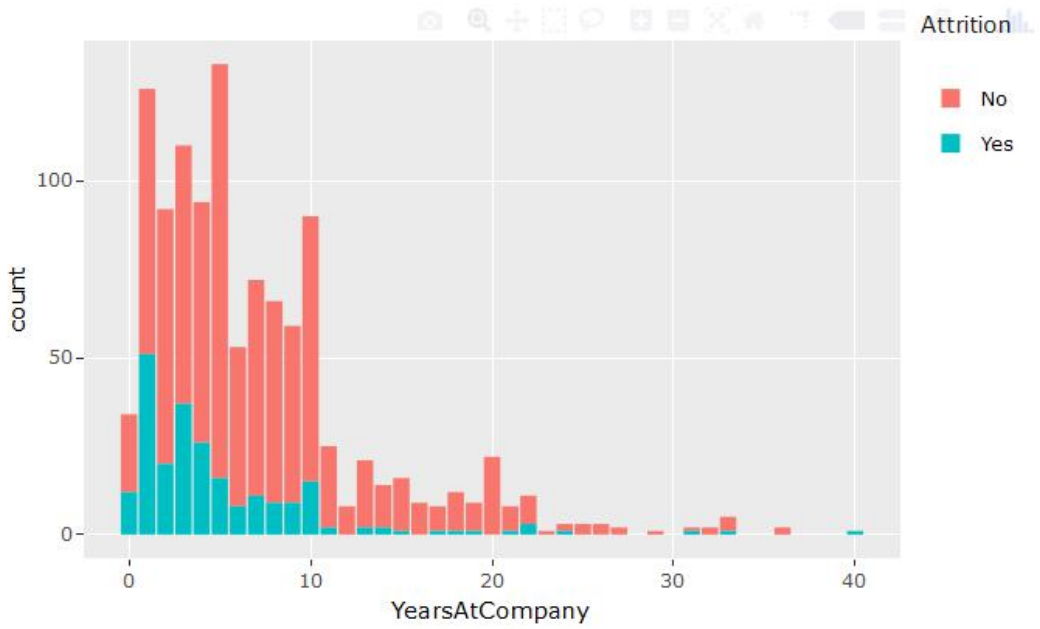


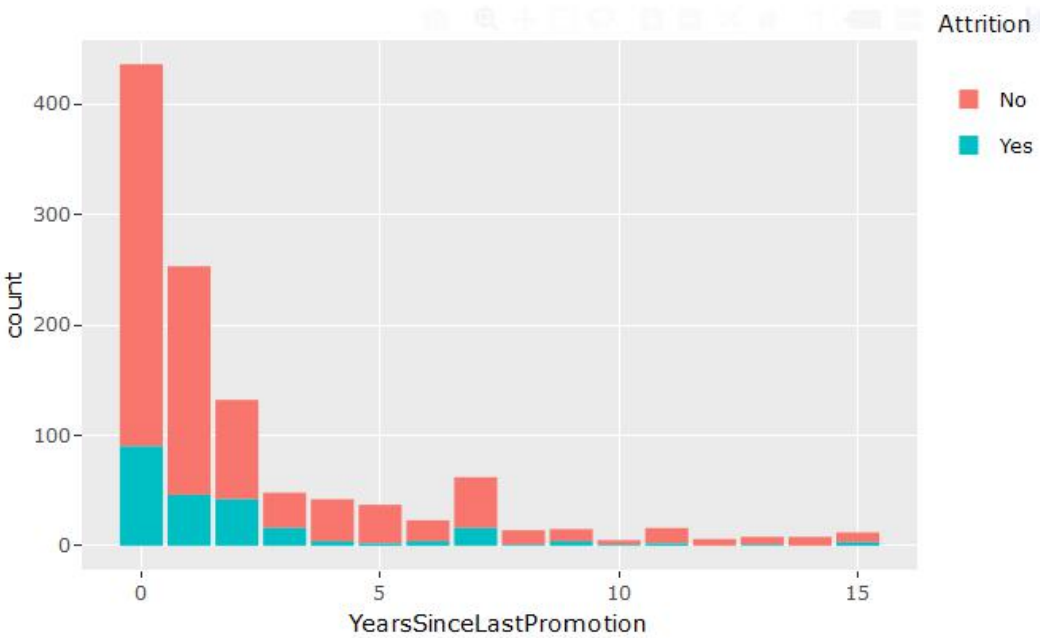
Fig Attrition by Monthly Income



Years at Company:

Larger proportion of new comers are quitting the organization. Which sidelines the recruitment efforts of the organization.

Fig Attrition by Years At Company



Years Since Last Promotion:

Larger proportion of people who have been promoted recently have quit the organization.

Fig Attrition by Years since last promotion

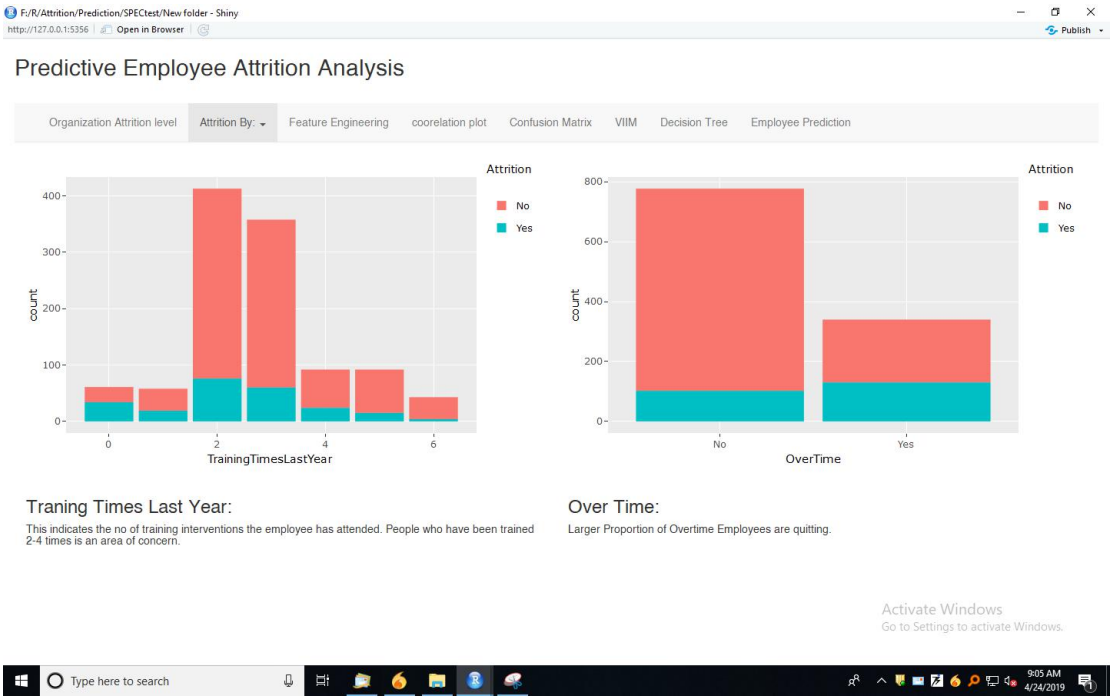
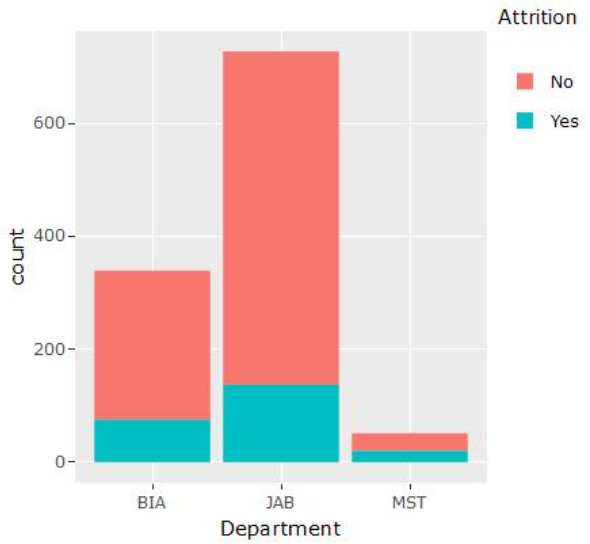


Fig Attrition by Training time and Over time



Department:

Among people attrited employees from HR dept. are less.It is because of low proportion of HR in the organization.

Fig Attrition by Department

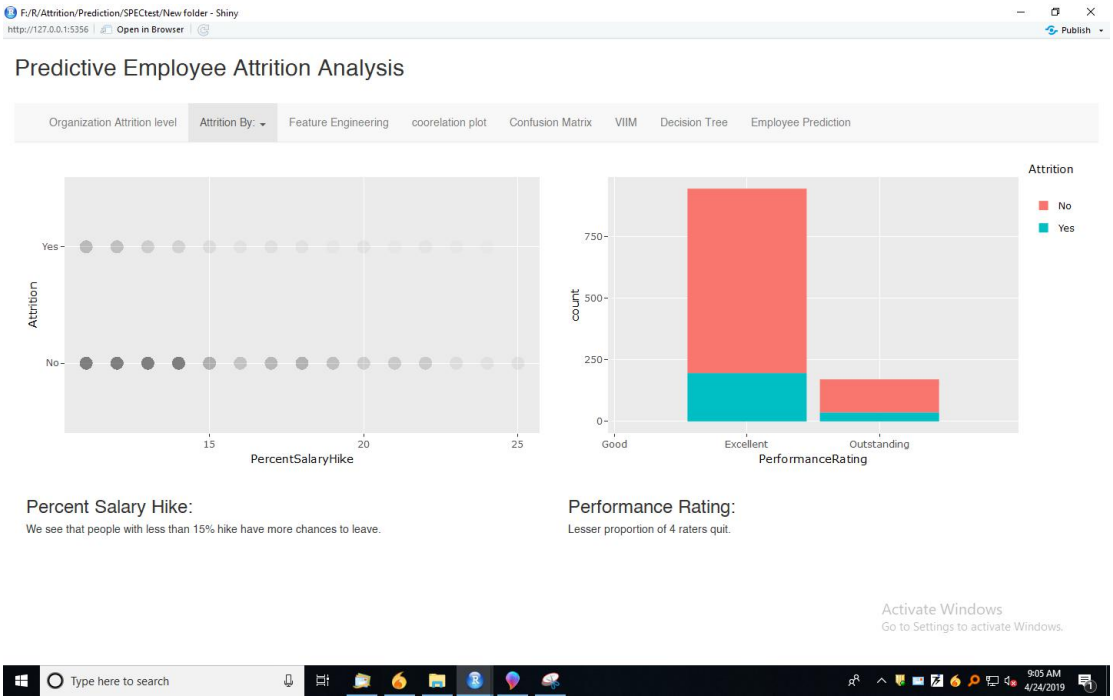


Fig Attrition by Percent salary hike and performance rating

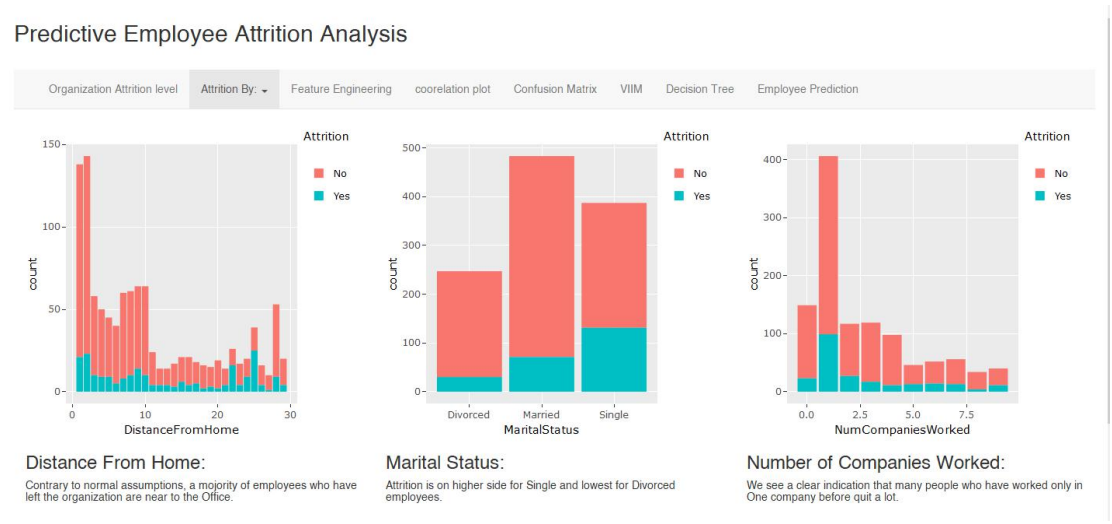


Fig Attrition by Distance from home, marital status and num. Of com. worked

Feature Engineering

Feature engineering is one aspect which provided a huge impact on the outcome rather than the model. Here, we try at creating new features with the existing variables we have based on my assumptions.

Tenure per job: Usually, people who have worked with many companies but for small periods at every organization tend to leave early as they always need a change of Organization to keep them going.

Years without Change: For any person, a change either in role or job level or responsibility is needed to keep the work exciting to continue. We create a variable to see how many years it has been for an employee without any sort of change using Promotion, Role and Job Change as a metric to cover different variants of change.

If we look at the plots in Fig 8, we see that there is an influence of these new features on the Attrition.

Compa Ratio: Compa Ratio is the ratio of the actual pay of an Employee to the midpoint of a salary range. The salary range can be that of his/her department or organization or role. The benchmark numbers can be a organization's pay or Industry average.

Here, we use the Comapny pay information to calculate our Compa Ratio at Department Level & Organiation Level.

People, with Compa Ratio less than 1, usually feel underpaid and show more tendency to leave the Organization in search of a better pay.

If we look at the figure (Fig 9), we can notice the effect of lower Compa Ratios.

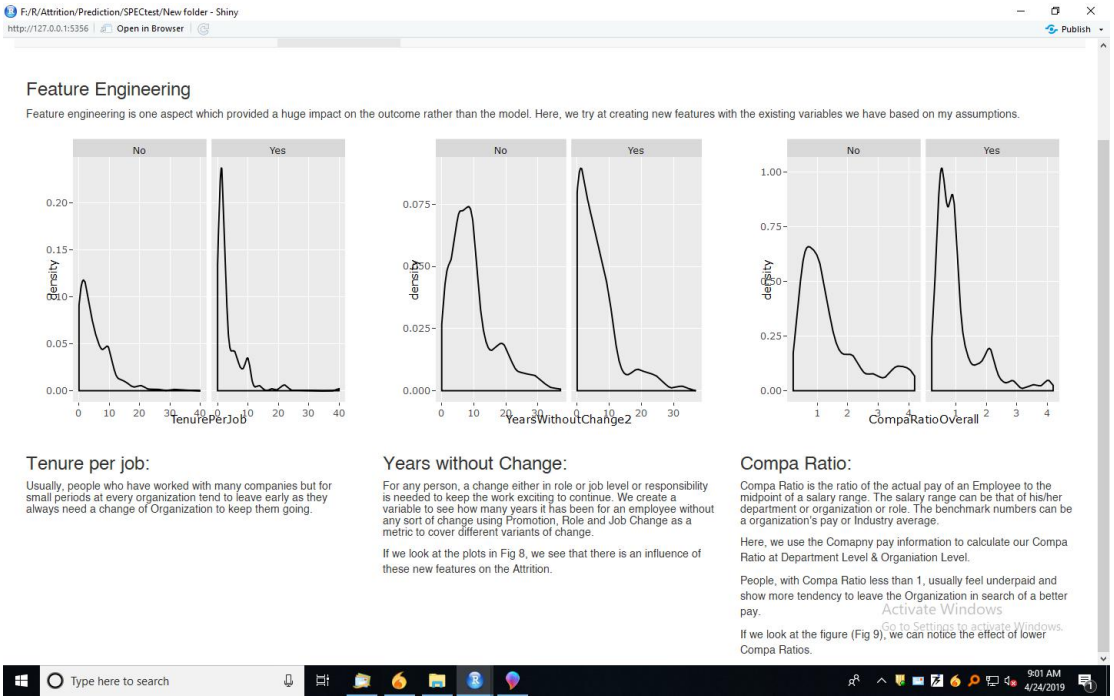


Fig Feature Engineering

Correlation of Variables

We see lot of correlation among the following variables

Years at Company, Years in Curr Role, Years with Curr Manager & Years Since Last Promotion - We will consider Total working years

Job Role, Department & Monthly Income - We will consider ‘Monthly income

Age is also one most correlated variable

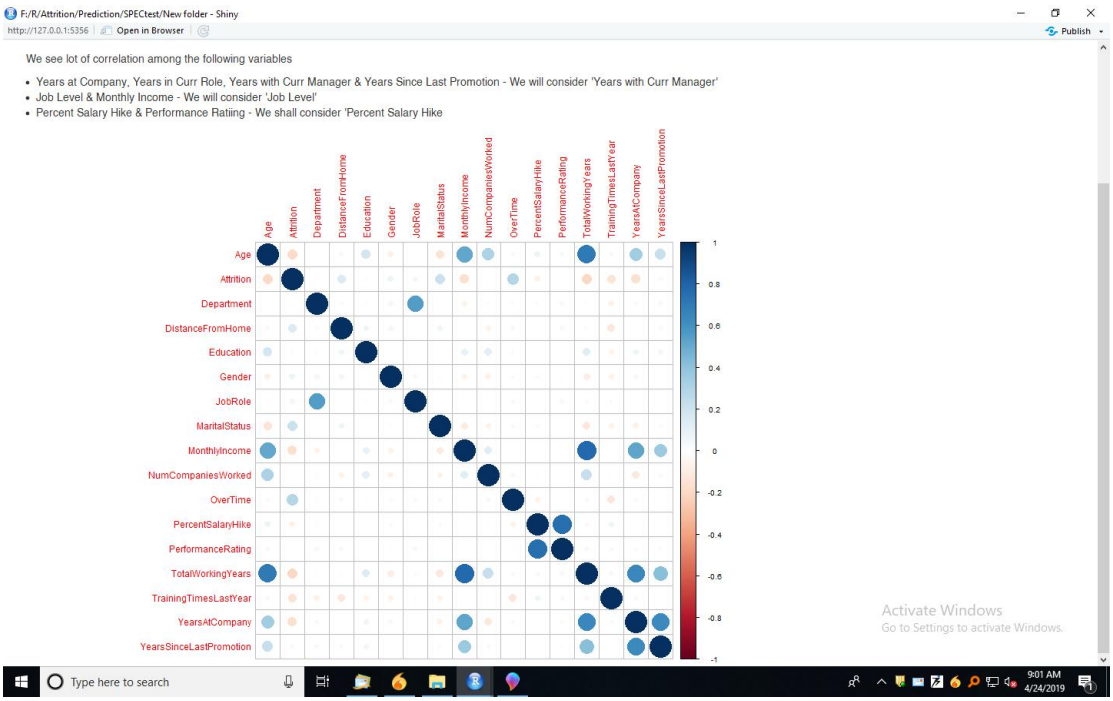


Fig Correlation plot

Basis of Selecting a Model

In order to find the best model we look at the confusion matrix comparing with the Actual Values.

Once we identify the best model, we try to tune it further to get the best results.

While looking at the confusion Matrix, if we consider only the Accuracy Values then our model is doomed to fail in Practicality.

For example, we get an accuracy of 90%,which looks like very good number, but we are able to correclty predict only 30% of the Minority Class (which is more important) and out of our Predictions only 30 % are correct then it is a challenge.

Out of overall 1600, if we have 300 people who are quitting and our model gives accuracy of 90 percent it is not enough to say that our model is good as the majority class is itself near 90% of the toal observations. Our aim is to predict the Minority. So, if our model is able to identify only 100 people (low specificity) and to identify that

100 people we are predicting in all 300 people (High Rate)- which means 200 are a wrong prediction.

In Practical Sense, if HR wants to talk to all the identified people they have to address atleast 300 employees to actually address the concerns of only 100 people which is also just 30% of overall Attrition.

Results

After checking all Models, we find that XGBTree (Boosted Decision Tree) works the best for us with a decent specificity rate (> 50%) and a very low error rate (< 30 %).

The overall accuracy is 89% which is also very good.

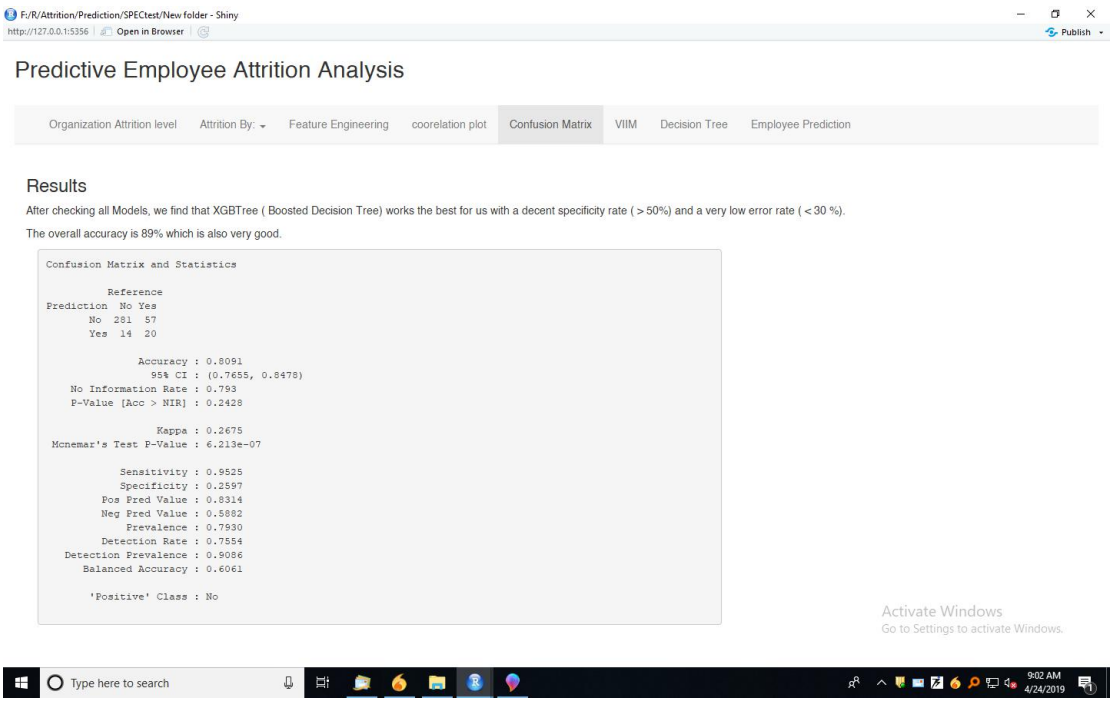


Fig Confusion Matrix

Variable Importance Plot

Dotchart of variable importance as measured by a Random Forest model. As we can see monthly income has highest importance in building our model and Performance rating don't has that much of importance in random forest.

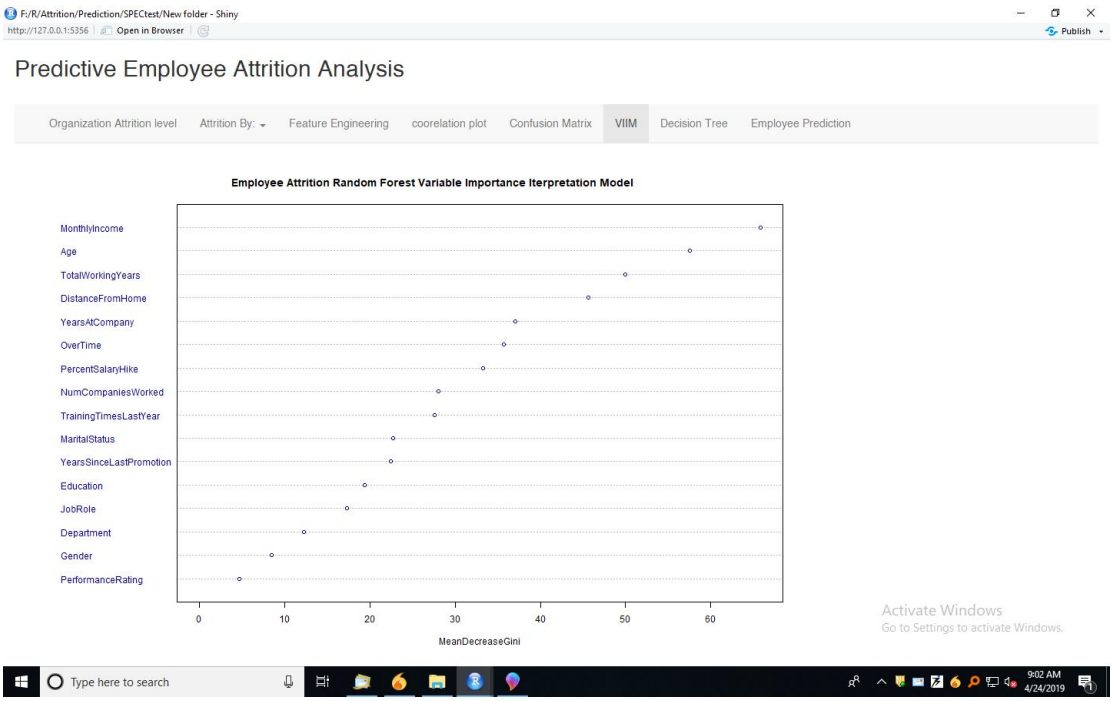


Fig Confusion Matrix

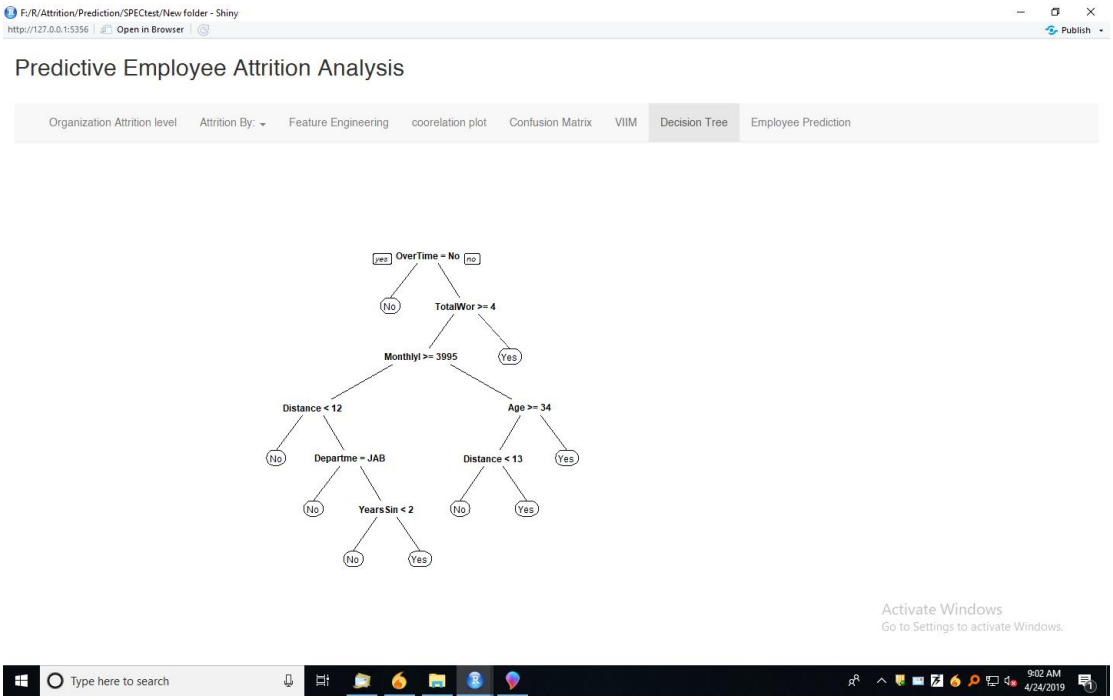


Fig Decision Tree

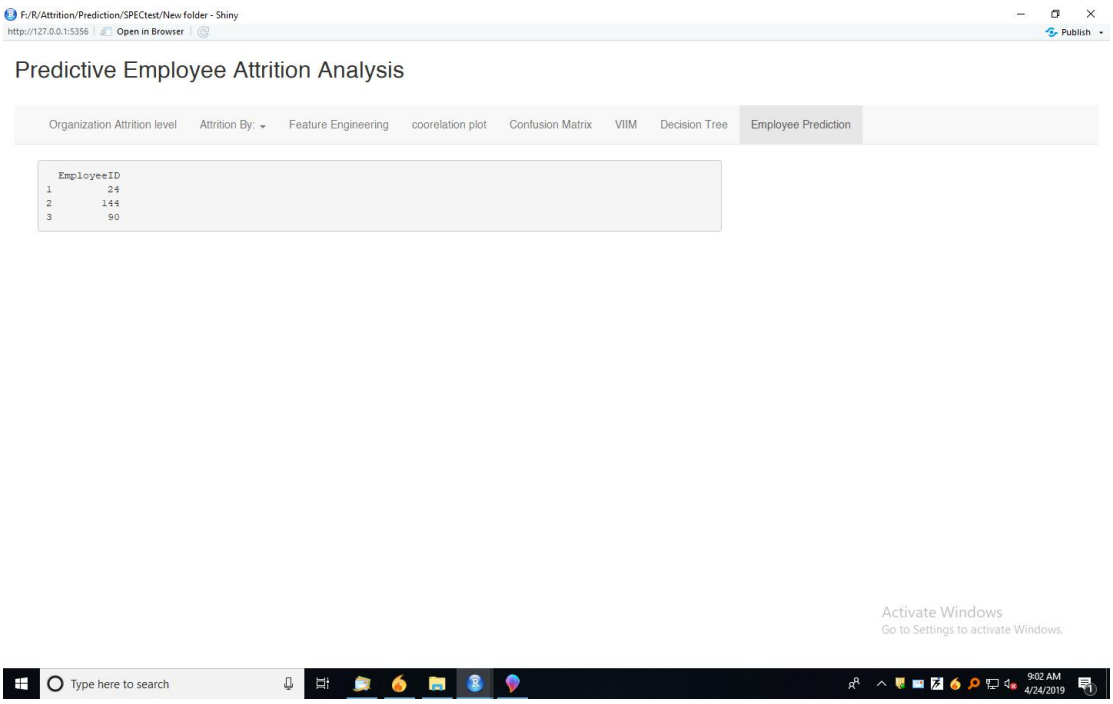


Fig Predicted Employee list

Conclusion

So, our best single model is XGBTree and further work can be done by looking at ensemble and stacking of the models which can help improve our metrics.