**HR Analytics: Employee Turnover Forecast**

**PIM 5604: Predictive Modeling**

**Prof. Jose Cruz**

**Dec 14, 2017**

**HR Analytics**

Human Resource (HR) analytics is about analyzing an organization’s people problem. It involves gathering HR data and analyzing them to create business decisions, helpful to improve the business processes.

**Business Problem**

An organization gets negatively impacted by huge turnover of employees, some of which are:

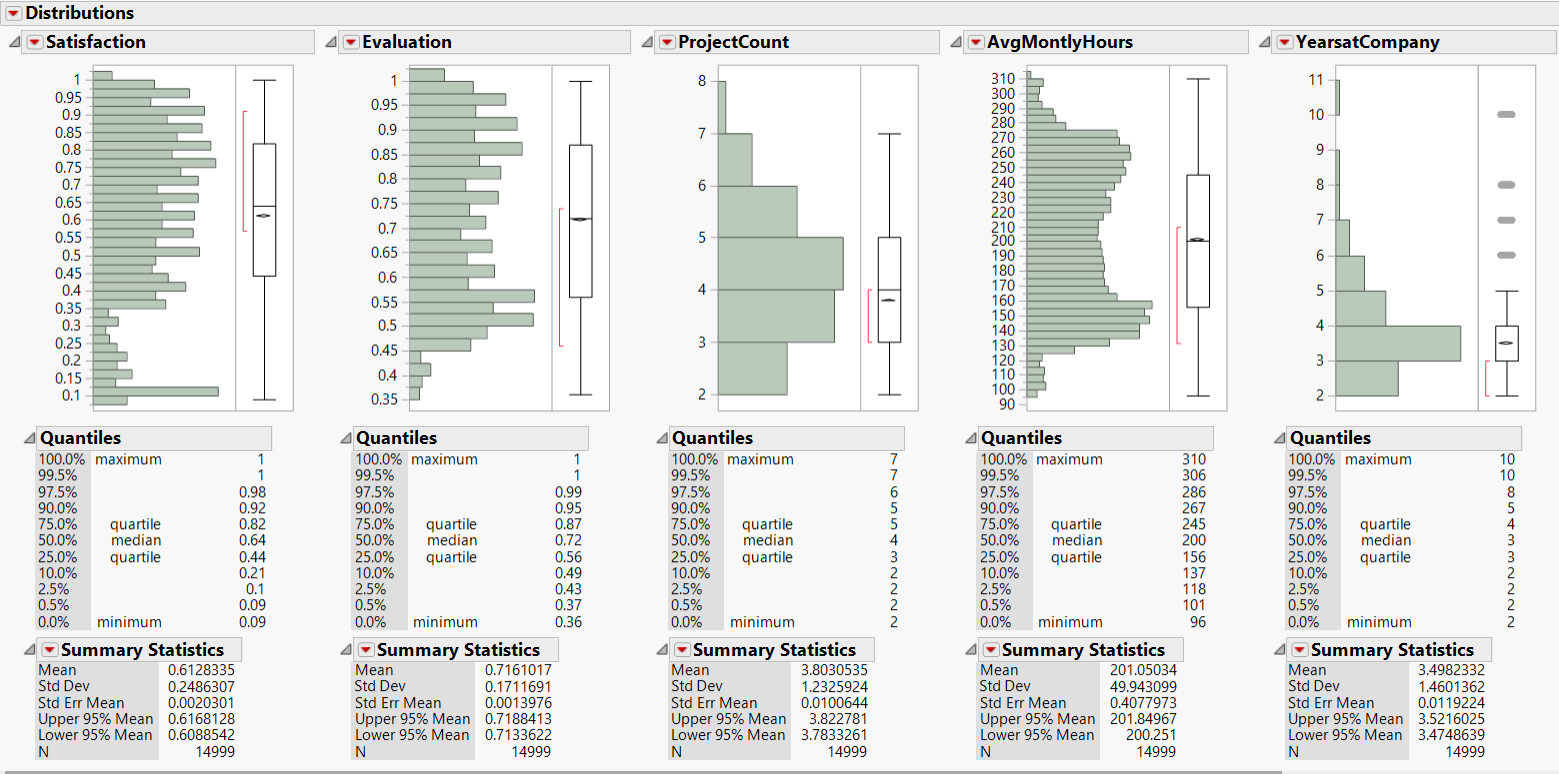
1. Revenue loss: The company must invest again for hiring, on-boarding and training the new resources which incurs added expenditure
2. Reduction in Marketing Return on Investment: Investors try to avoid investing in companies who do not have a stable workforce.
3. Lower Knowledge Base: New employees who join the organization are most of the time not completely aware of the strategy and policies of the company which reduces their productivity and ultimately affects the performance of the company.

Due to the high impact of huge turnovers on an organization, it is important for the company to address the reasons because of which employees tend to leave. This project aims at observing the relation among the reasons of turnover and predicting the potential employee who will be leaving next. The project provides recommendations to the organization which should be taken care of in order to reduce the turnover rate.

**Data Description**

The data is obtained from Kaggle website and is available to download from: <https://www.kaggle.com/ludobenistant/hr-analytics-1>**.** The data has 15000 entries where our target variable is Turnover, which describes whether an employee has left the company or not. We have various predictors in the data like Satisfaction Level, Last evaluation result, Number of projects assigned, Average monthly work hours, Years with the company, whether they have had a work accident, whether they have had a promotion in the last 5 years, the department, and their salary level.

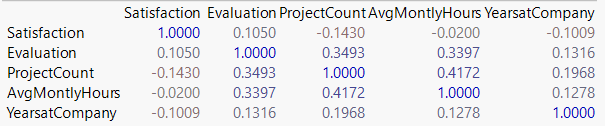
**Data Preprocessing:**

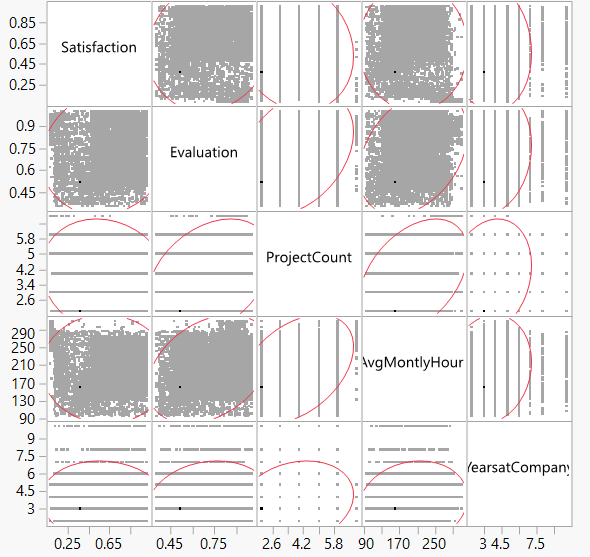
****

As seen above, the data doesn’t have outliers. The only probable outliers are in the category of Years at Company, which we chose not to classify as outliers, because there will be employees who have work experience at the company of more than 5 years, so we cannot rule them out from our analysis.

Also, we found that the data is free of missing values, so there was no need to impute or delete any row, and we proceed with the original data for our analysis.

**Finding Patters: The Correlation Matrix:**



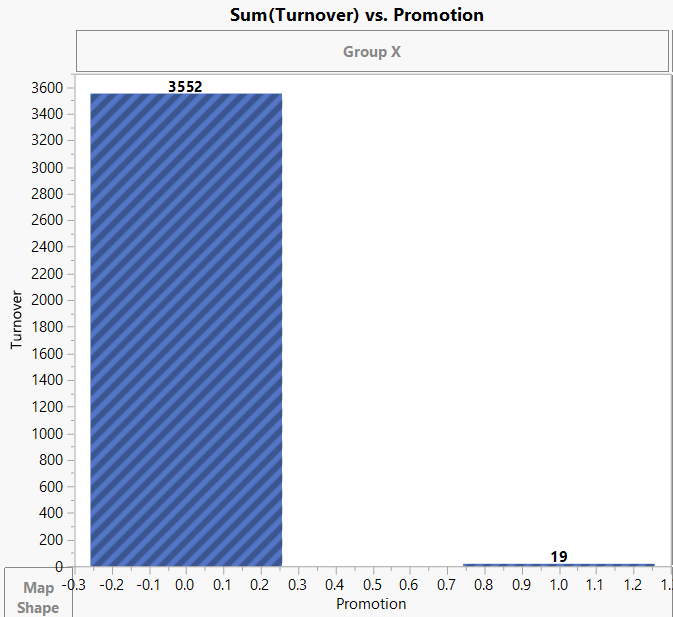
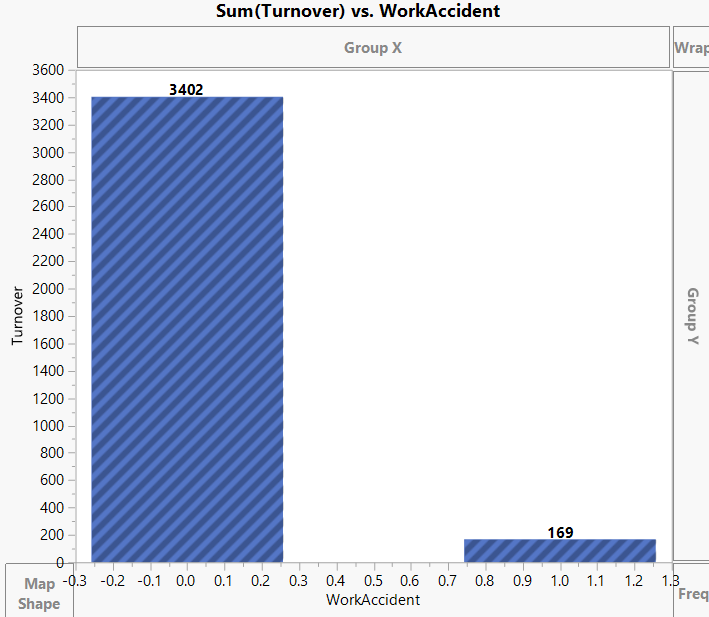


From the above figure we see that we do not have any pair of variable which show any considerable relationship pattern with each other. The maximum correlation found was between Project Count and No of Monthly Hours, of 0.4172, which we did not take into our consideration.

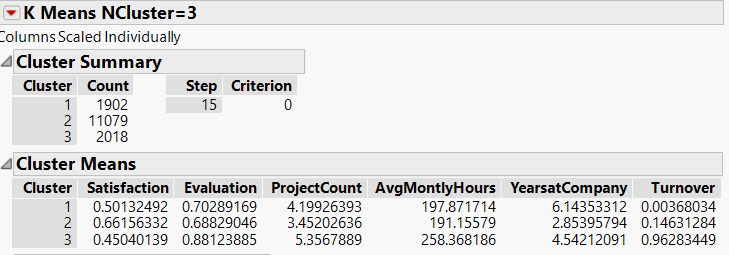
So, from this analysis we were unable to classify any variable as insignificant.

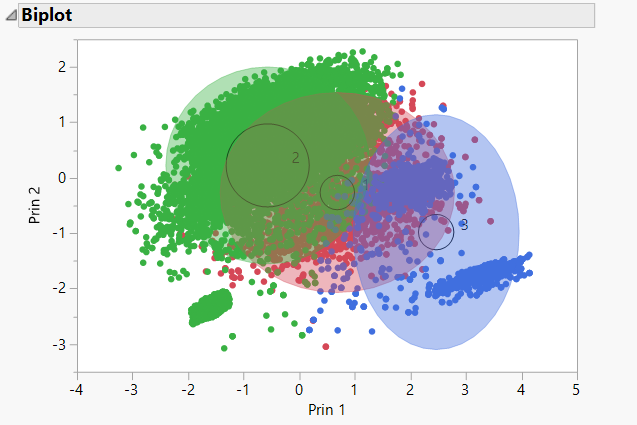
**Clustering**

**Why Employees are leaving company?** Selecting data of employees who left the company.

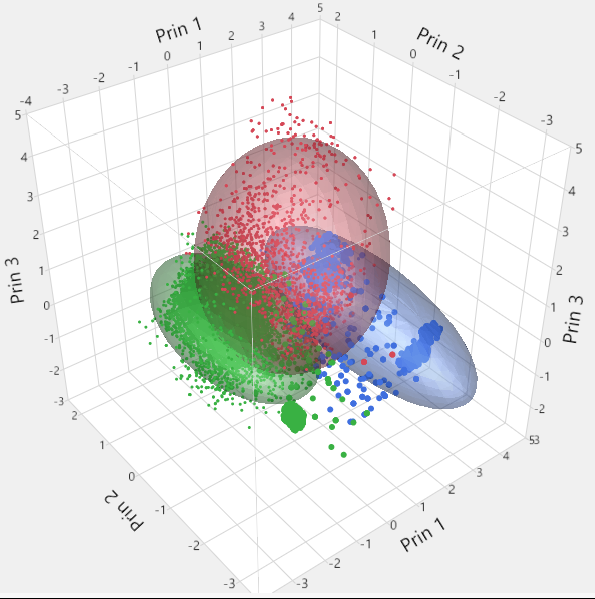
 

Promotion seems to be significant for modeling whereas WorkAccident does not show clear picture.





Satisfaction, Evaluation, AvgMonthlyHours, YearsatCompany and ProjectCount seem to be significant in clustering.



Cluster 1

Employees with low satisfaction, good evaluation, 198 average work hours and almost 6 years at company constitutes lowest fraction of employees leaving the company.

Cluster 2

Employees with good satisfaction, good evaluation, 191 average work hours and almost 3 years at company constitutes low fraction of employees leaving the company.

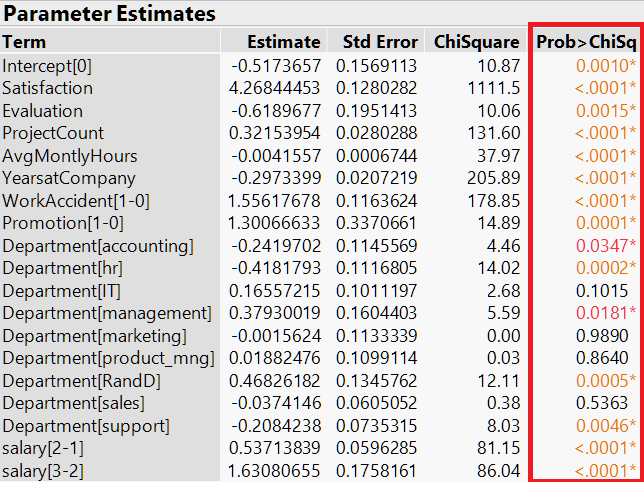
Cluster 3

Employees with low satisfaction, high evaluation, average work hours more than 250 and almost 4.5 years at company constitutes **HIGHEST** fraction (0.96) of employees leaving the company.

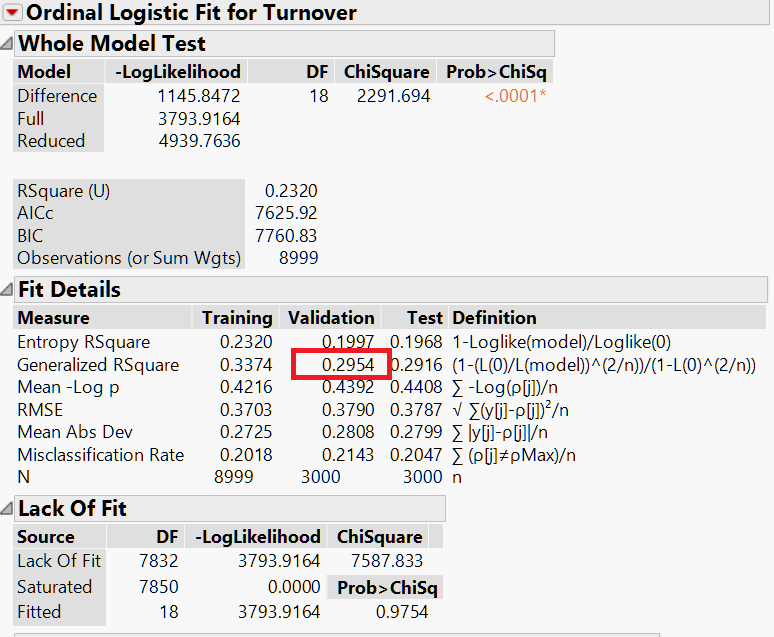
**Modeling**

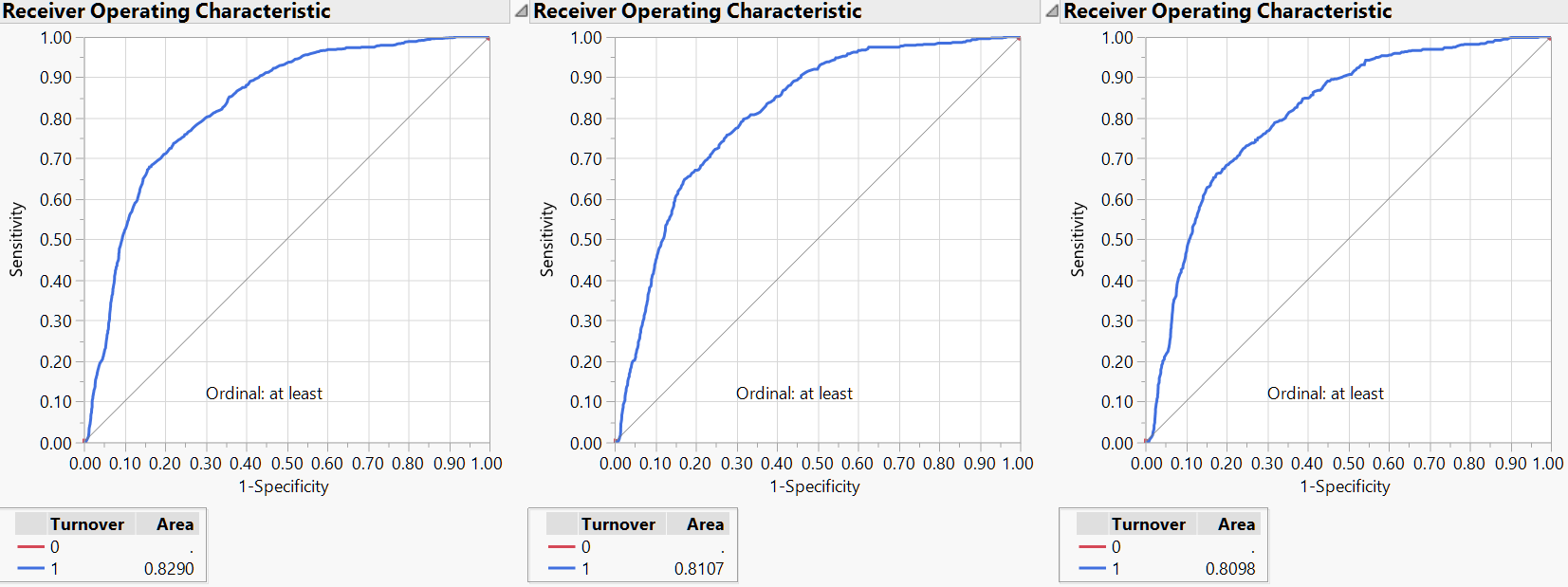
**Logistic Regression**

First, we ran a regression model using all the variables:



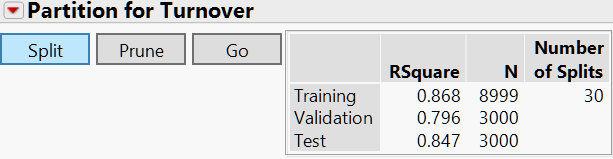
It shows that all variables play important roles in the modeling. Based on these estimates, we can get some insights about the potential effects of these factors on whether an employee decides to leave the company.

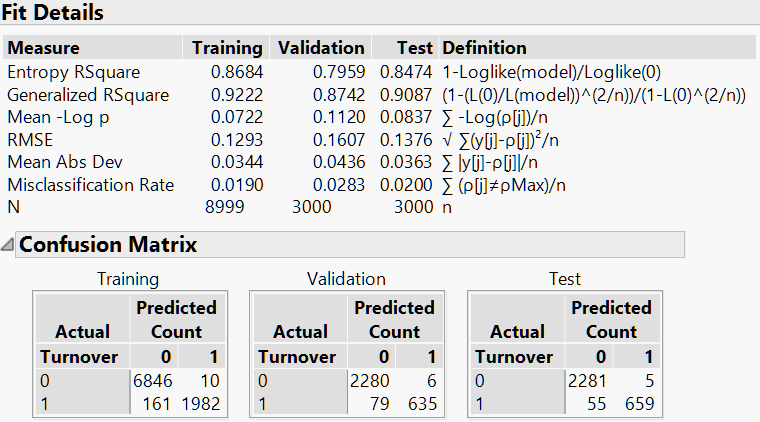


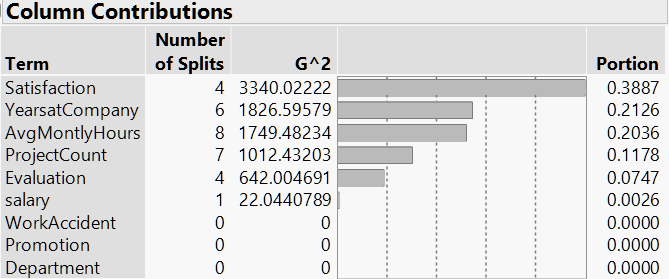
These performance measures clearly state that the model is poor in terms of predicting the target variable. To better predict if an employee would likely leave, we would develop some other kinds of models.

**Decision Tree**

Then we ran a decision tree model using all the variables:

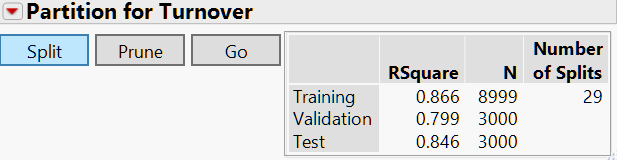


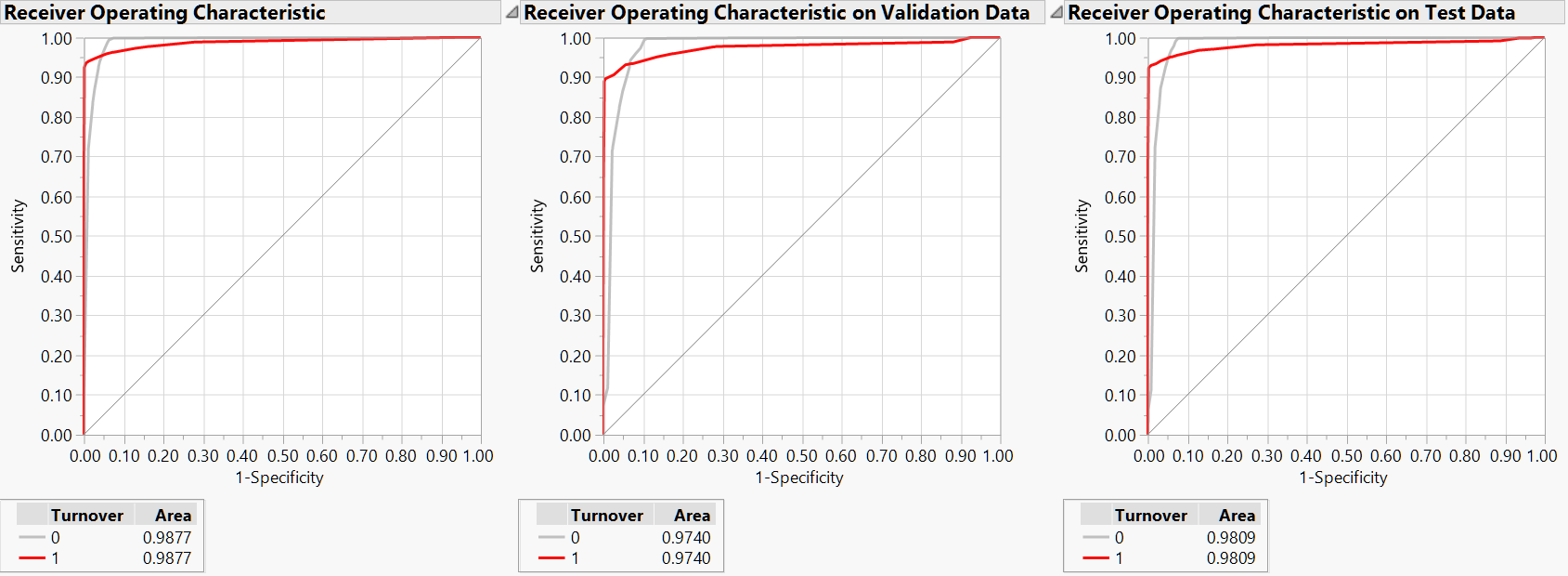
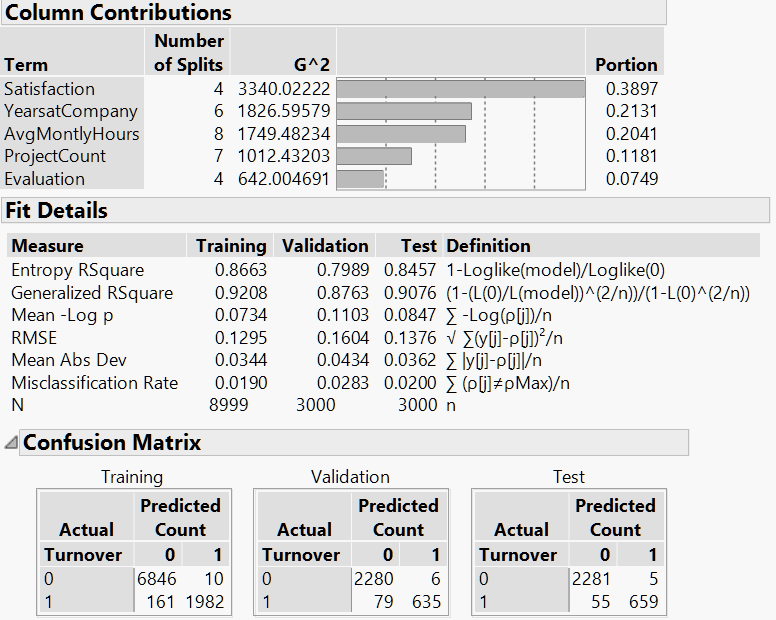


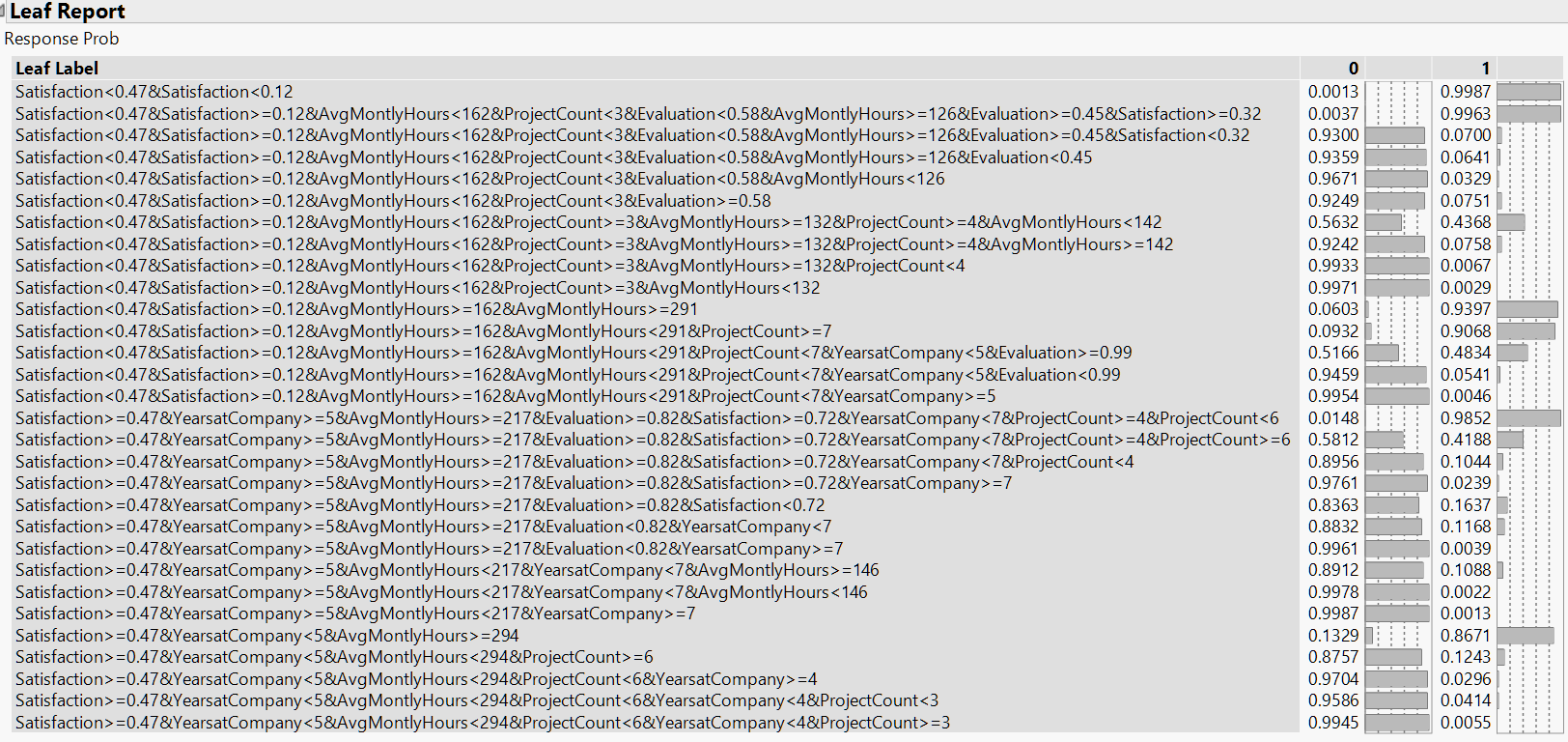


As shown above, the variables *Salary, Promotions, WorkAccident, and Department* play no significant part in the Random Forest model. Therefore, we just got rid of these variables and reran the model.

Here’s the result of the model including only the significant variables.

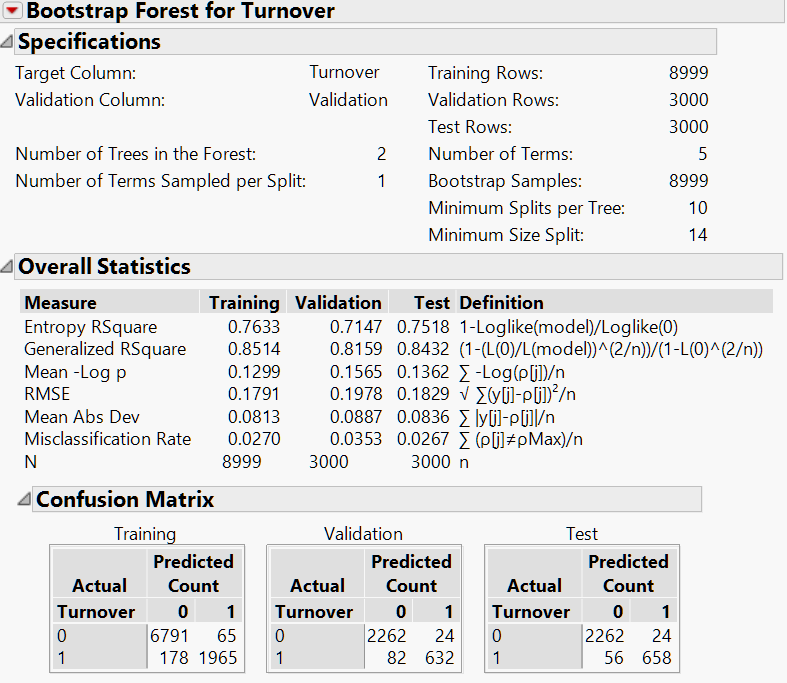


As you can see, this model performs slightly better than the regression model.

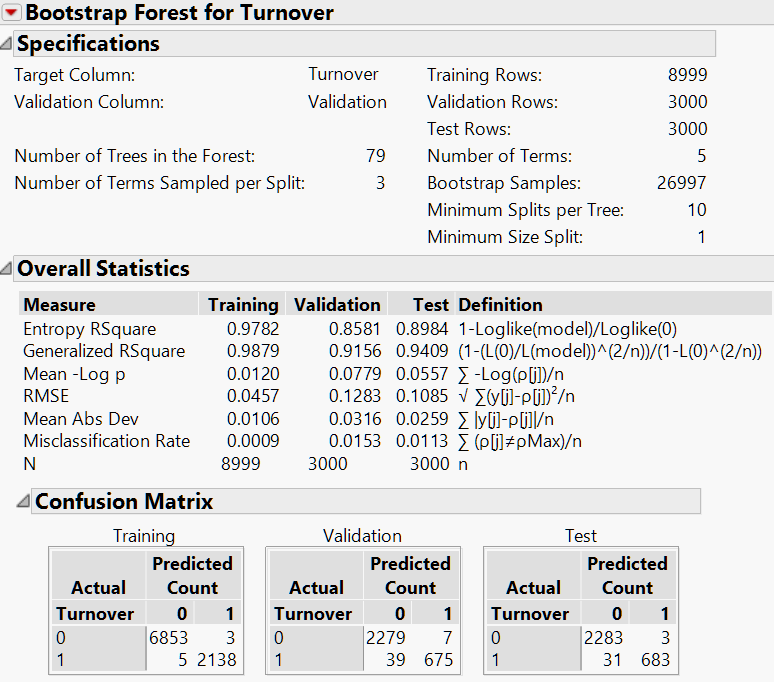


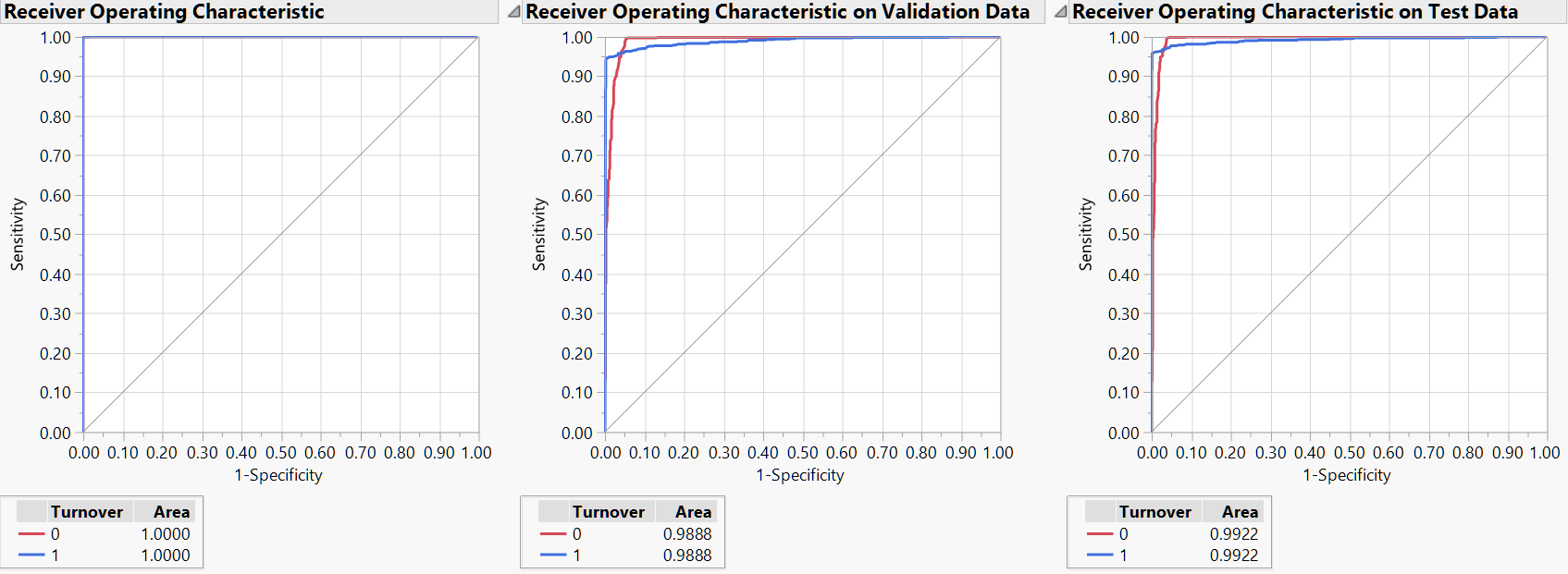
**Bootstrap Forest**

To further improve our tree model, we ran a bootstrap forest model using the same variables:



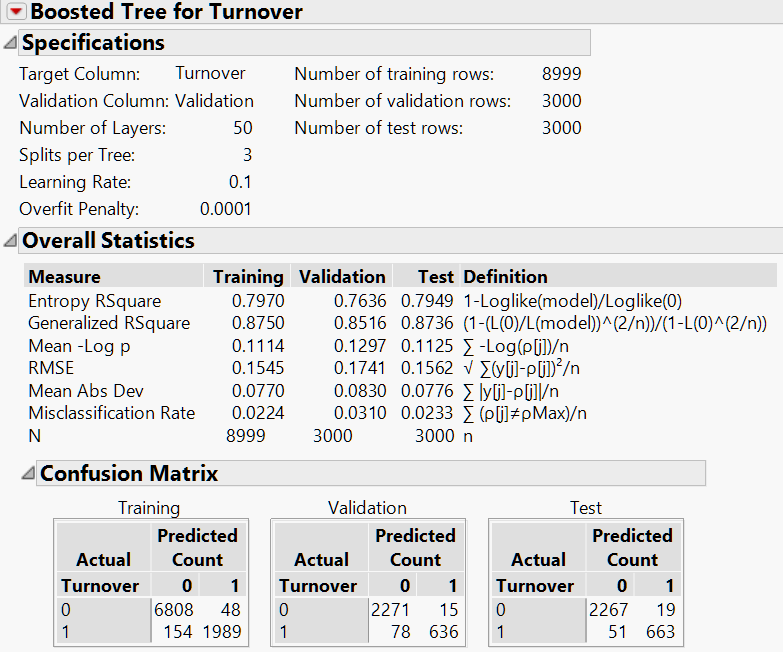
The above model is the best one I got after running several times. However, this model is still worse than the decision tree model we just got, possibly because of the default settings we used. Hence, we tried to tune the model to develop a better model.



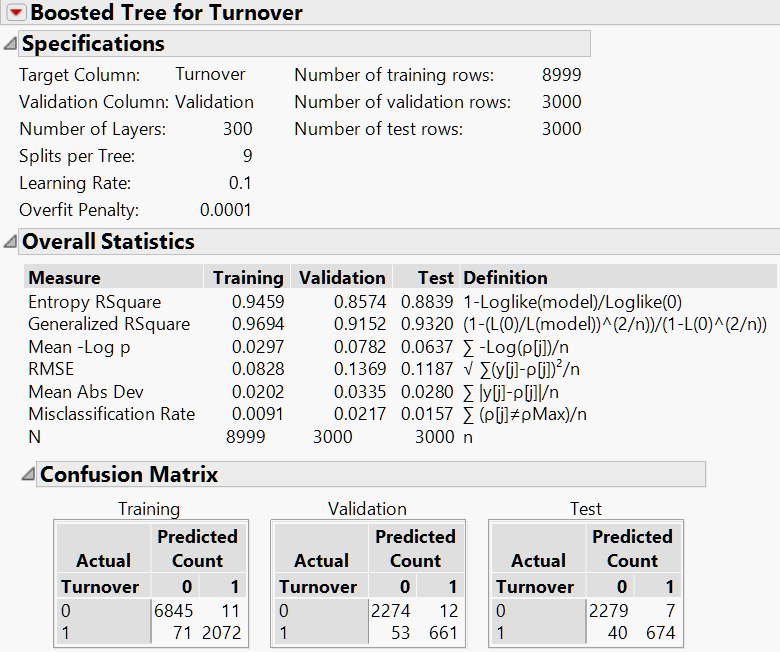


**Boosted Tree**

Later, we tried a boosted tree model using the same variables:

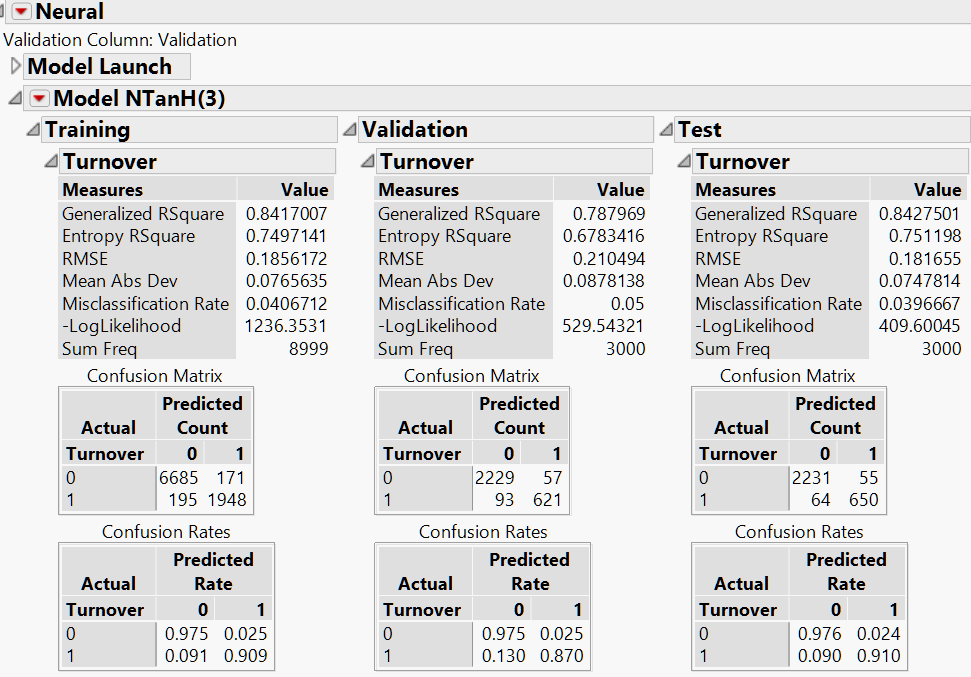


Like our bootstrap forest model, the boosted tree model with the default settings could be improved by tuning.

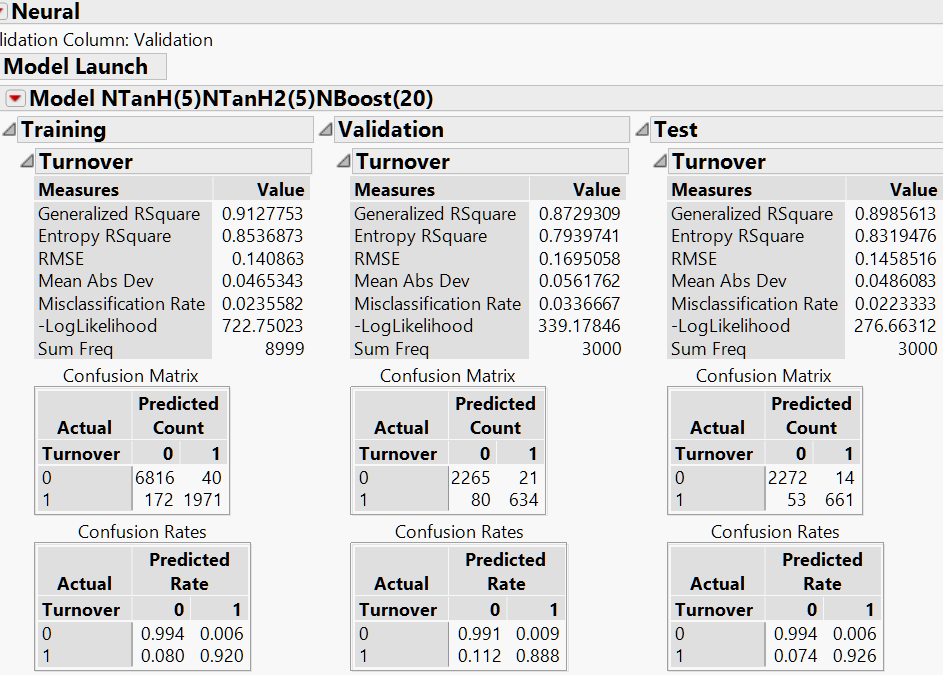


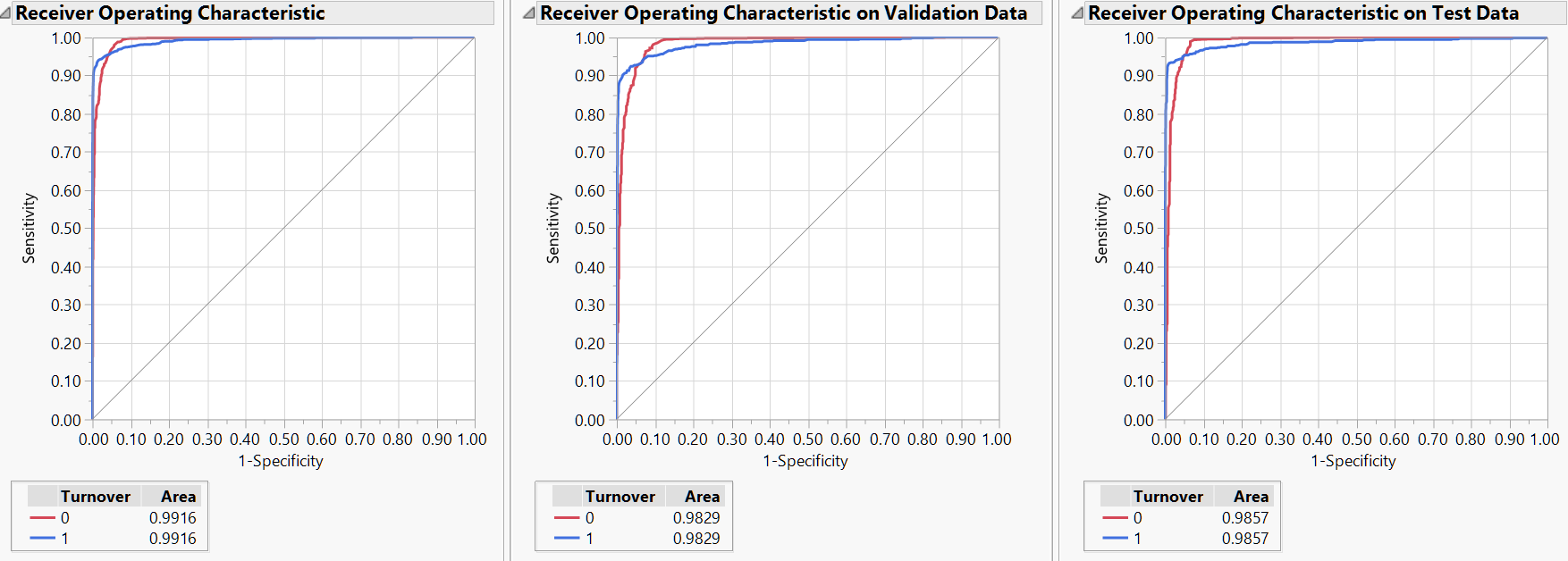
**Neural Network**

We ran a neural network model using all the variables:

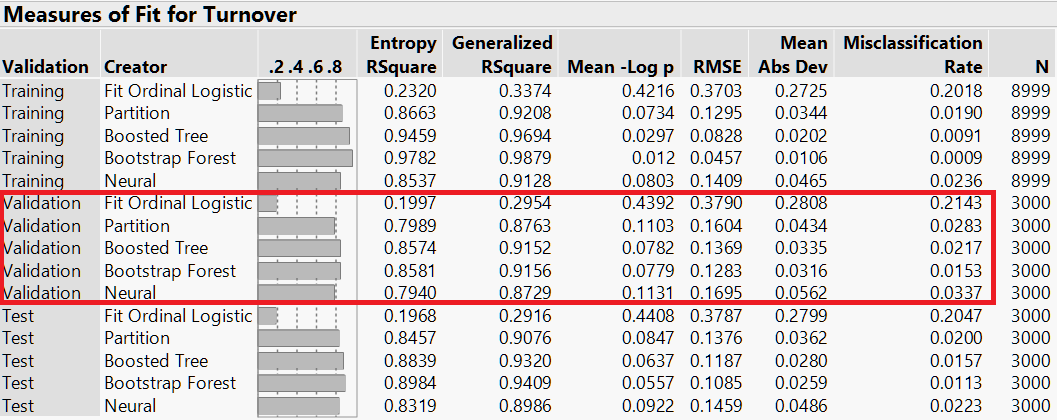


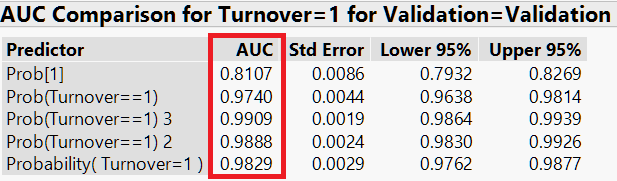
Using the default settings, this model isn’t as good as the previous ones we built. With some pruning and removing the variables *Salary, Promotions, WorkAccident, and Department*, we got a better model.



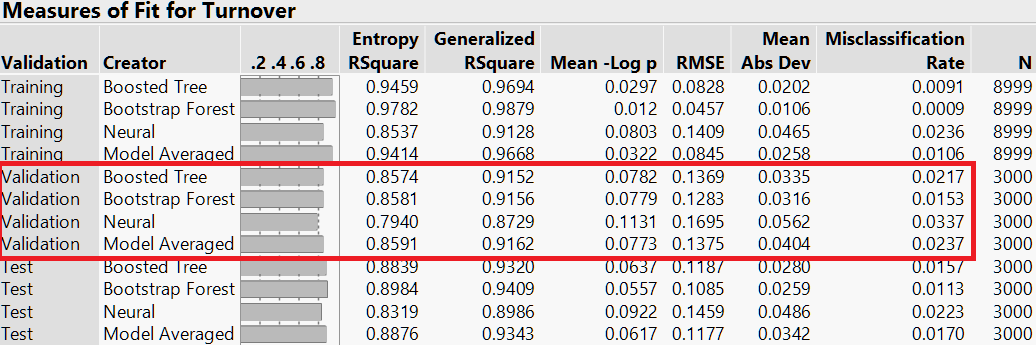


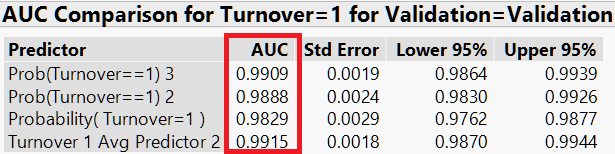
**Model Comparison**





Our Bootstrap Forest model and Boosted Tree model are the best 2 models with almost the same level of accuracy. Which one should we choose? Neither. Instead, we combined the Bootstrap Forest model and the Boosted Tree model with Neural Network model to make an ensemble model with better performance.

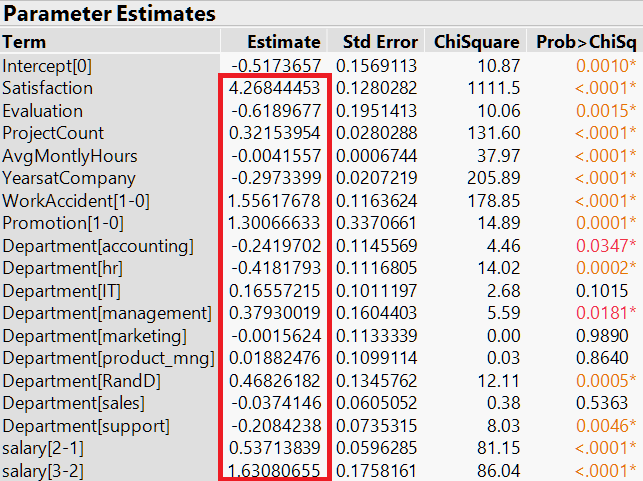




As you can see in the above comparison, the ensemble model is our chosen model.

**CONCLUSION**

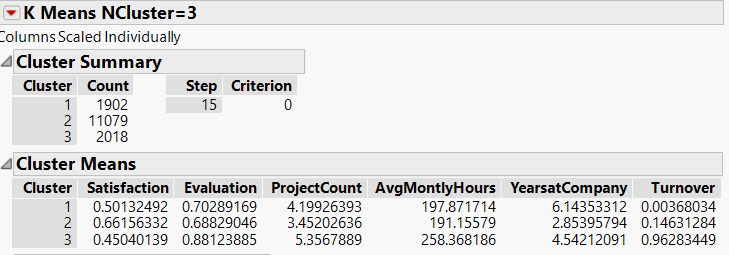
On one hand, based on our logistic regression model, we could interpret its implications on potential contributors to the possibility that an employee leaves the company.



As shown above, satisfaction level, projects, work accident, promotion, and salary seem positively correlated with the likelihood of an employee leaving the company, while evaluation result, working hours, and work years seem negatively correlated with the likelihood of an employee leaving the company. Additionally, certainly departments have higher turnover rates than others.

This analysis raises a few questions:

1. It makes sense that work accident can probably leads an employee to leave and more projects assigned can make one too overwhelmed to stay. But why do employees with higher satisfaction, recent promotion, and higher salary also tend to leave the company? One explanation is that such an employee may find better career opportunities outside the current company with better income and position. It can also be that some employees may not be honest answering the satisfaction survey.
2. It makes sense that employees with better evaluation result find it better to stay in the current company for its career and employees who have stayed in the company long enough are more used to the environment here. Otherwise, they have had left the company. But why do more working hours also lead an employee to stay at the company? One explanation is that working hours don’t necessarily indicate the level of workload. After all, there are not very extreme values in terms of work hours.



Based on our clustering analysis, the vast majority of leaving employees are those who have been with the company for around 4.5 years, have been given the highest evaluation, have worked for most hours and projects, and have willingly expressed their dissatisfaction.

Therefore, the company may need to consider decreasing the workload for certain employees, especially those outstanding ones. Moreover, promotion or higher salary don’t necessarily help keep an employee as she may still try to find better job opportunities with lower workload or better salary and position. Hence, the company may come up with other ways to maintain employees, for example, building nicer corporate culture or offering better company amenities such as training programs, more holidays, and flexible schedule. By such initiatives, the company can fundamentally reduce its turnover rate in the long run.

On the other hand, our ensemble model has remarkable accuracy in terms of predicting if an employee would leave the company. It means that the company can likely intervene before an employee decides to quit the job, thus bringing in considerable economic and strategic benefits by keeping its experienced and talented human resources.