Human Resource Management Analytics

Navaneesh Gangala
Data Scientist
https://www.linkedin.com/in/navaneesh/

Problem statement:

- The main goal here is to predict whether an employee will stay or leave within company.

- Besides this,
  - Key drivers for attrition are to be identified
  - Employees shall be classified into High ,Medium and Low risk profiles in terms of attrition
  - Predicting thresholds for key drivers of attrition
  - Employee level risk analysis showing supporting and contradicting features

Approach

- The dependent or target variable is binary in nature i.e. Terminated or not Terminated
- In the given data, the target variable “Terminated” (0 = stay, 1 = leave) shall be predicted using the other features.
- Hence classification could be the solution.
- To check which of the classification algorithm suits to the situation ROC - AUC Analysis has been done
- From the model selected, the following have been derived

    Key drivers that influence the employee Termination,

    Thresholds for each driver

    Various segments of risk profile has been derived.

Data Preparation

- Initial Exploration :
  - Total : 15 Variables
  - Target : Terminated
  - Unique Key : Employee_code (can be used to get the risk profile of the person once the model is ready)
  - Remaining 13 predictors are identified as
    - Numerical : Tenure, TimeLastPos, LastRating, Annual.Income, Year.of.Birth
    - Categorical : Rehire,Department,Job.Level,Has.been.promoted,Client.work.travel,Education,Gender,Marital.Status

- Anomaly Detection :
  - Structure and factor levels are good
  - Tenure - Negative values detected and removed.
  - Employee_code – few records found not to be unique hence removed
  - LastRating - Median(2) can be used to replace imputed mean (1.95) to have better results
  - Education - Two levels with similar meaning made into one(Bachelor's Degree → Bachelors Degree)
  - Last Time Promoted values < Tenure - Records where Validation failed were removed

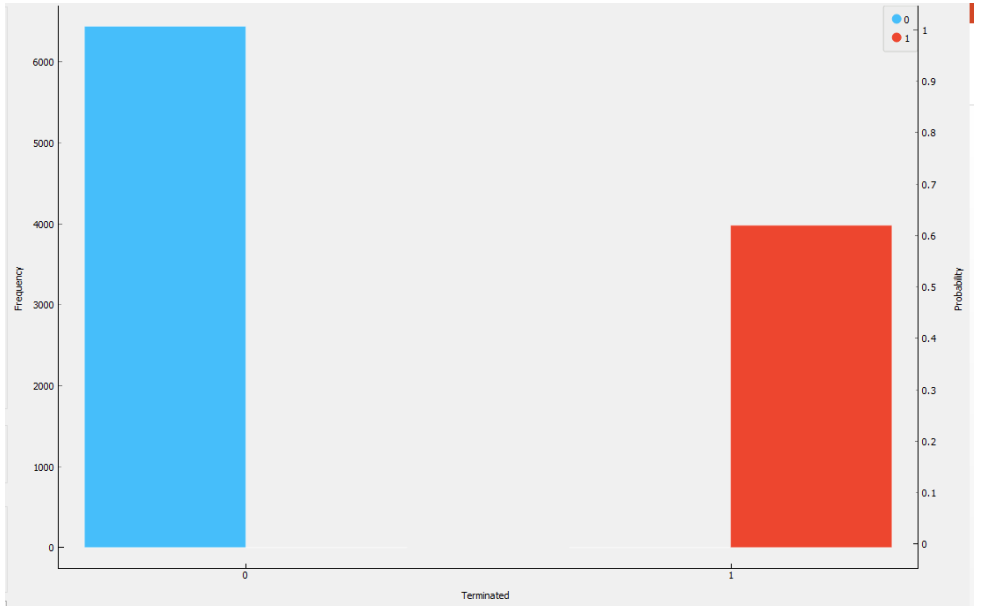
Cleaned Data : HR SAMPLE V2.csv

Rehire	Terminated	Employee_code	Department	Job Level	Tenure	TimeLastPos	Has been promoted	LastRating	Client work travel	Education	Gender	Marital Status	Annual Income	Year of Birth
FALSE	1	133715	Finance	Lead Analyst	16	16	No	2	Medium Travel	Msc Analytics	M	Married	6638	1971
FALSE	1	91202	Audit	Senior Analyst	19	19	No	2	Medium Travel	Msc Analytics	M	Married	1004	1972
FALSE	1	51471	Audit	Staff I	33	33	No	2	Medium Travel	Msc Analytics	M	Married	643	1989
FALSE	1	95874	Consulting	Manager	39	39	No	2	Medium Travel	Msc Analytics	M	Married	5769	1964
FALSE	1	105869	Finance	Analyst	44	44	No	2	Medium Travel	Msc Analytics	M	Married	1043	1981
FALSE	1	20105	Tax	Staff II	79	79	No	2	Medium Travel	Msc Analytics	M	Married	1552	1963
FALSE	1	31135	Audit	Staff II	82	82	No	2	Medium Travel	Msc Analytics	M	Married	2884	1984
FALSE	1	52977	Risk Management	Lead Analyst	85	85	No	2	Medium Travel	Msc Analytics	M	Married	6828	1964
FALSE	1	115564	Risk Management	Analyst	89	89	No	2	Medium Travel	Msc Analytics	M	Married	2705	1988

Target variable

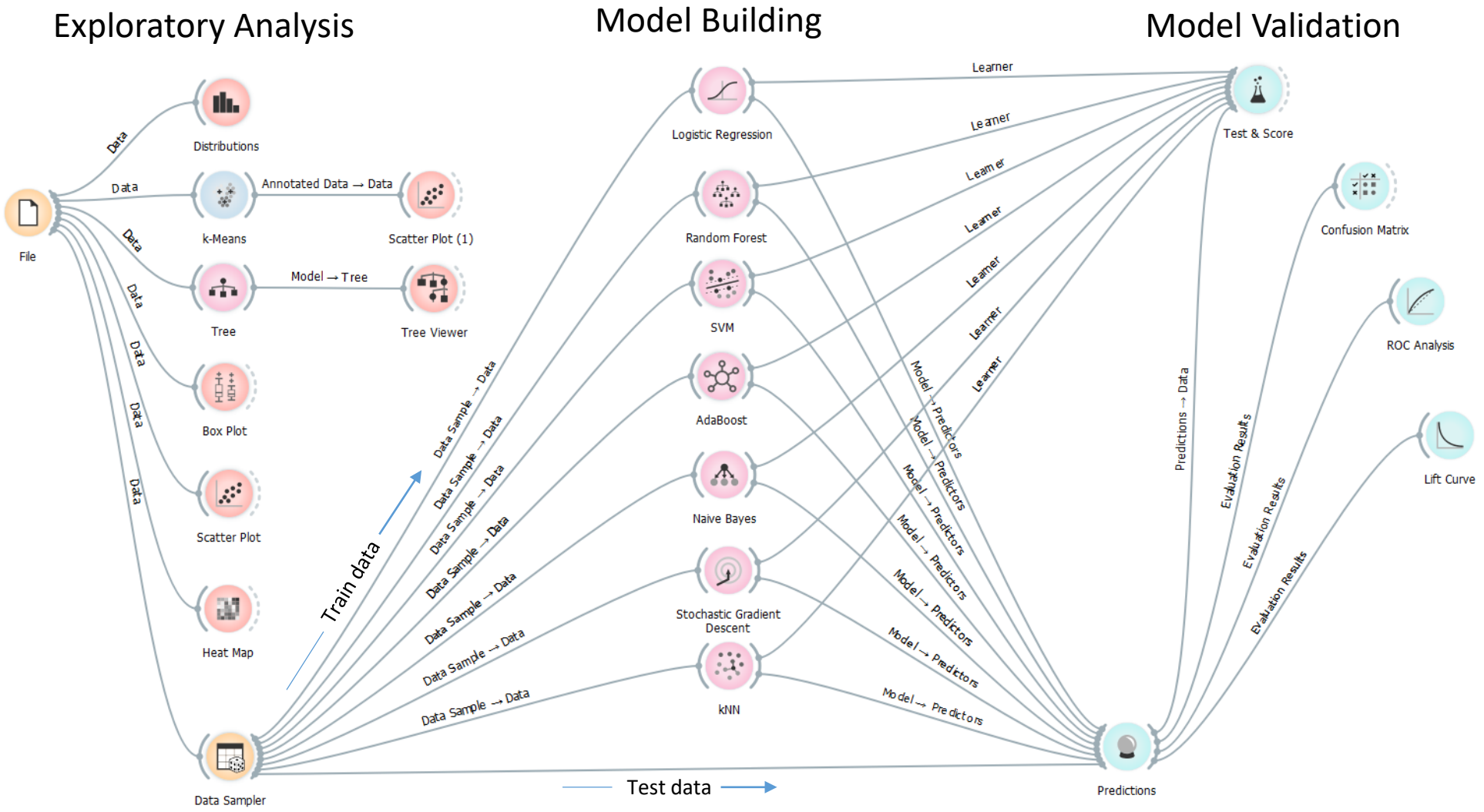
Unique key

```
   0    1
6439 3981
```

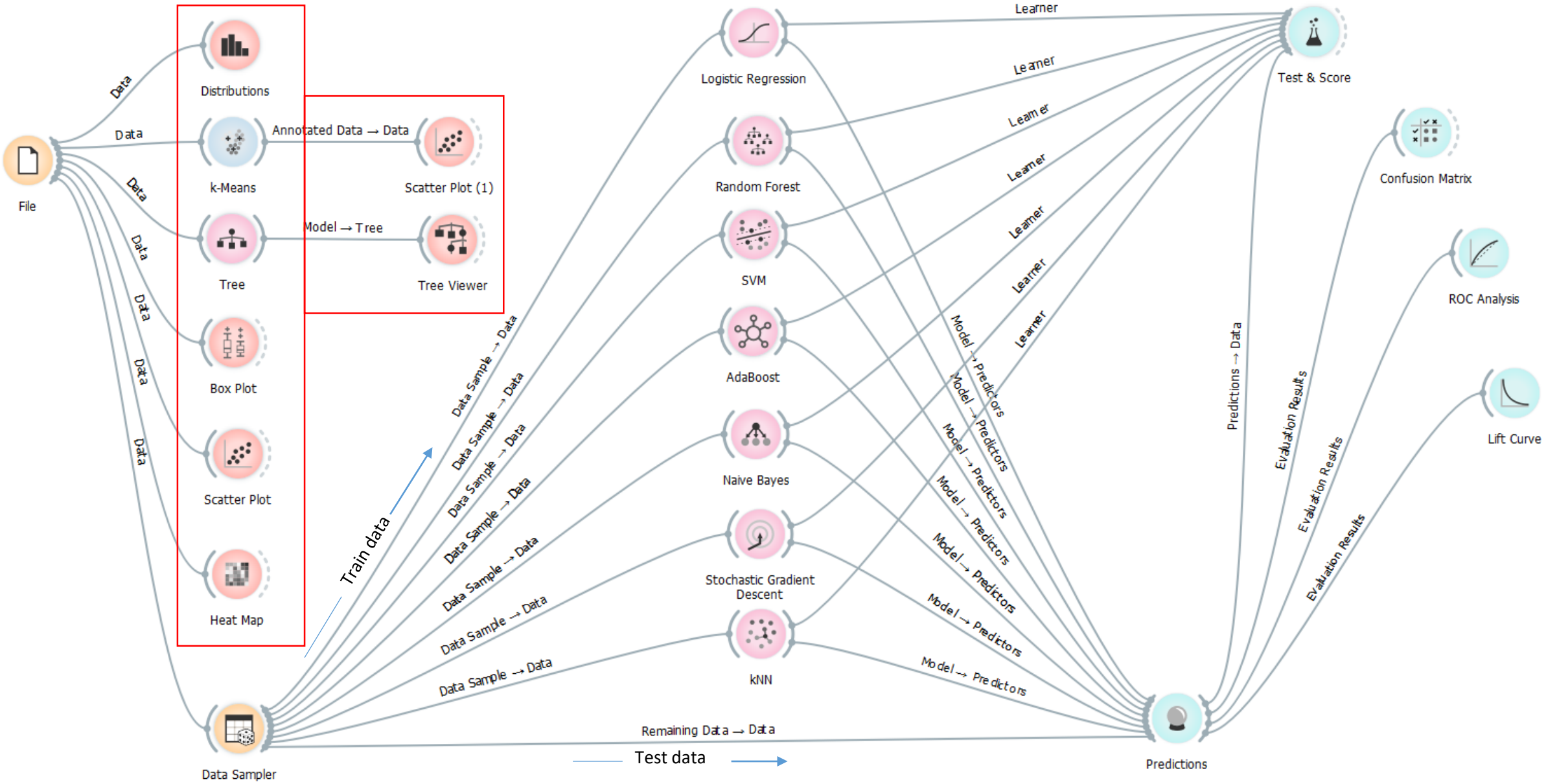


- Imbalance Problem has not being detected

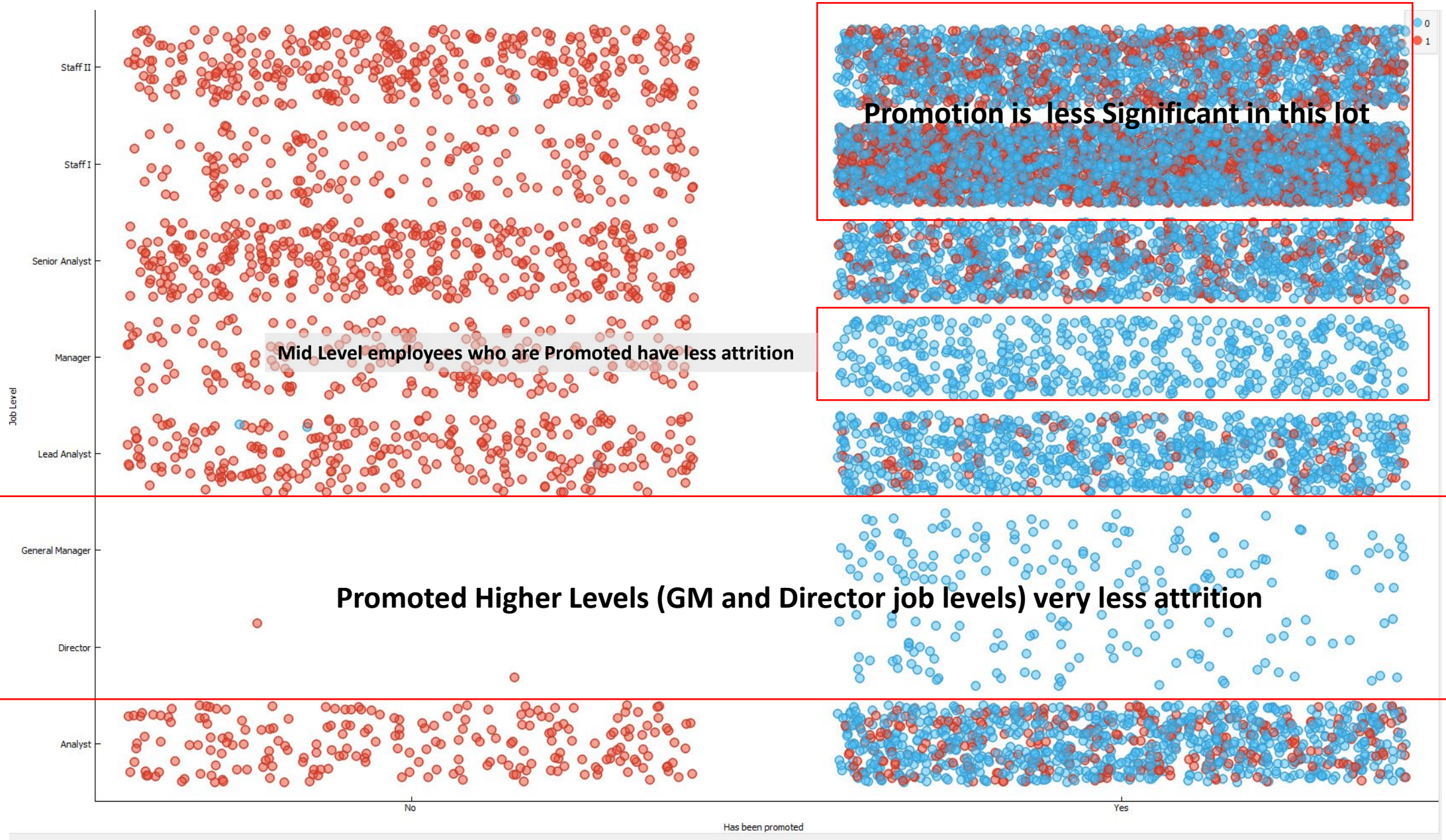
Project : Work flow



1. Exploratory Analysis of Data



Promotion can be very significant

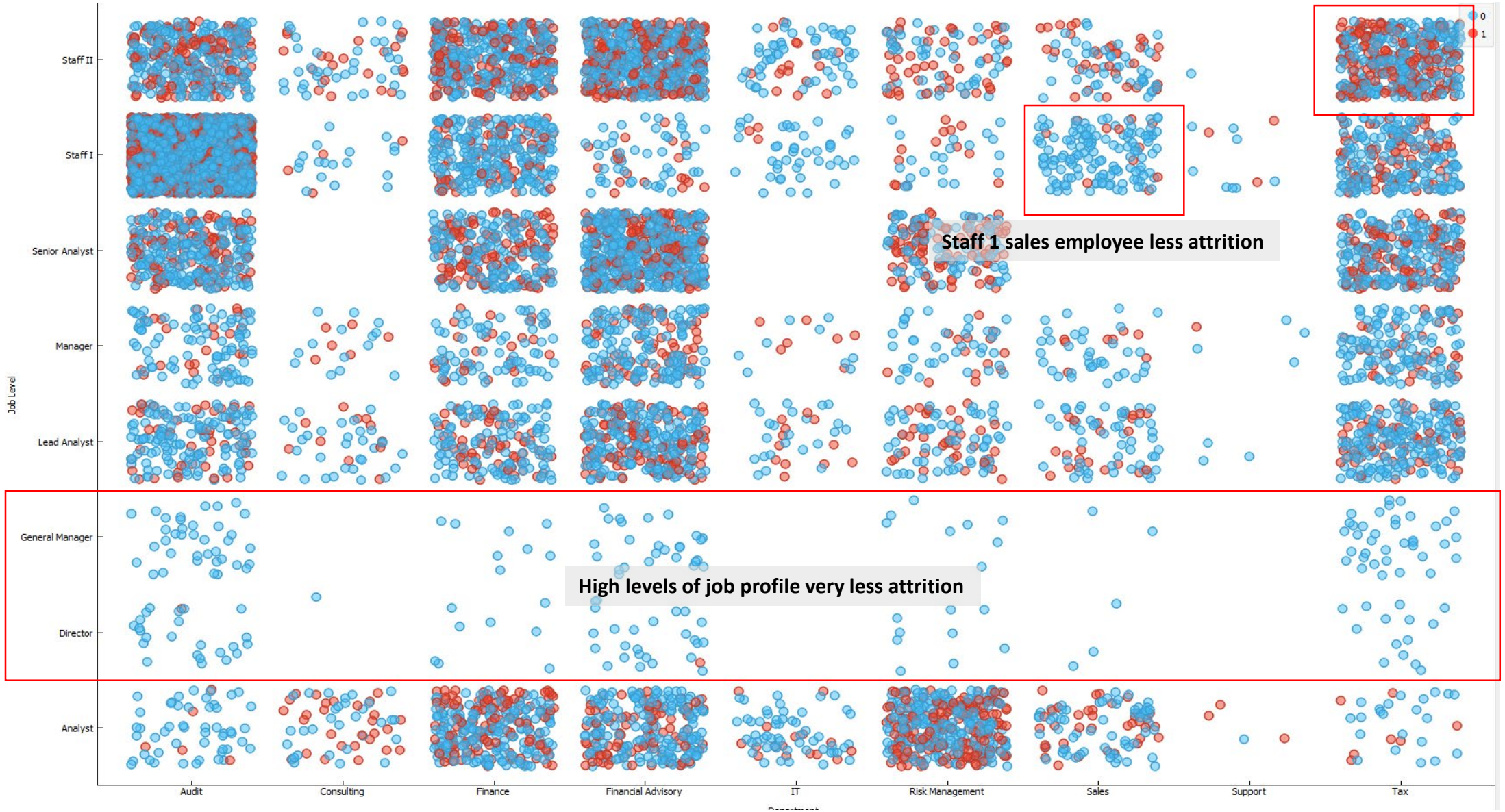


Job Level – With Department

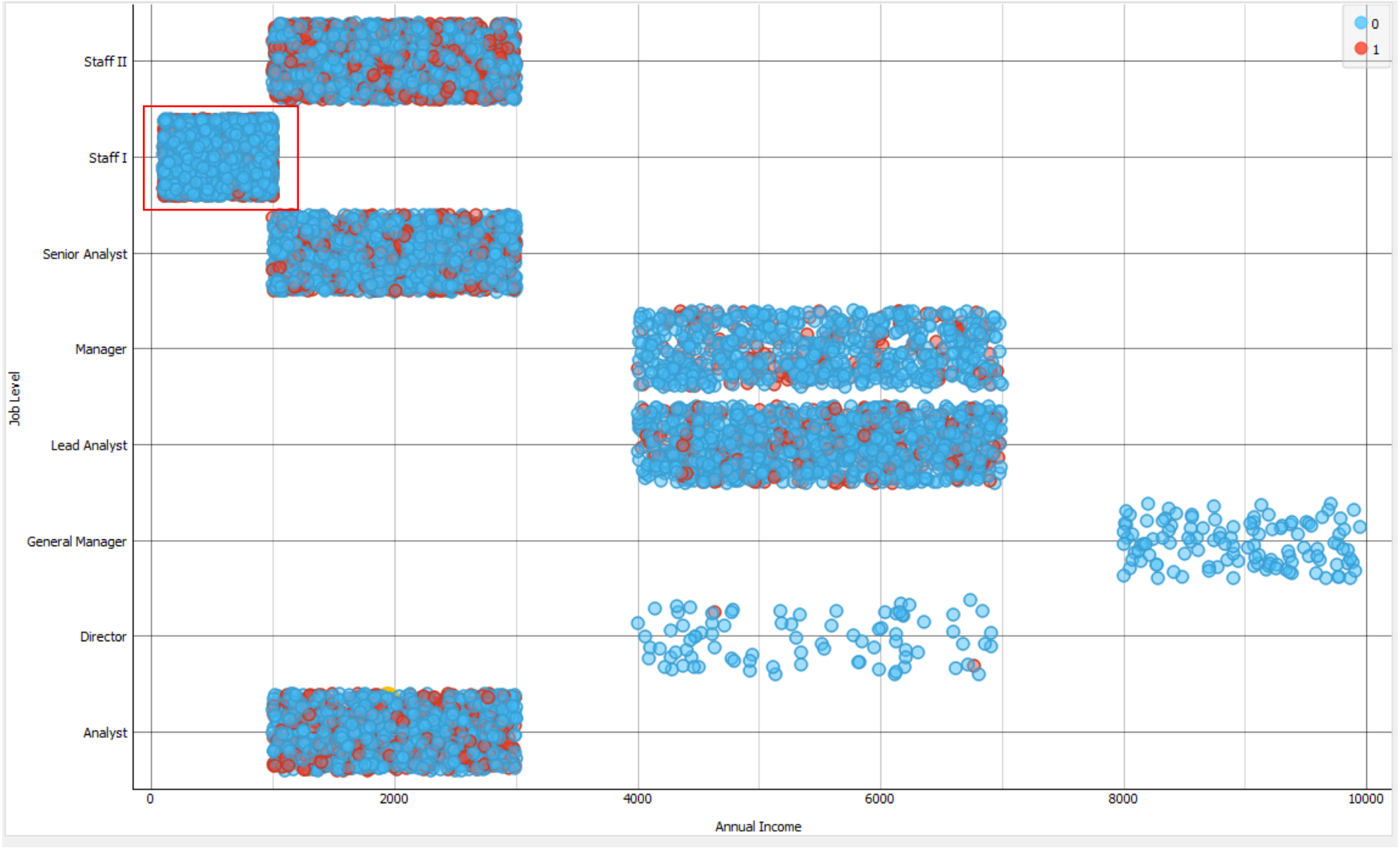
Job levels has significance but not as great as Promotion (High level jobs – less exits)

More terminations can be seen in high population hence model should reveal more about this

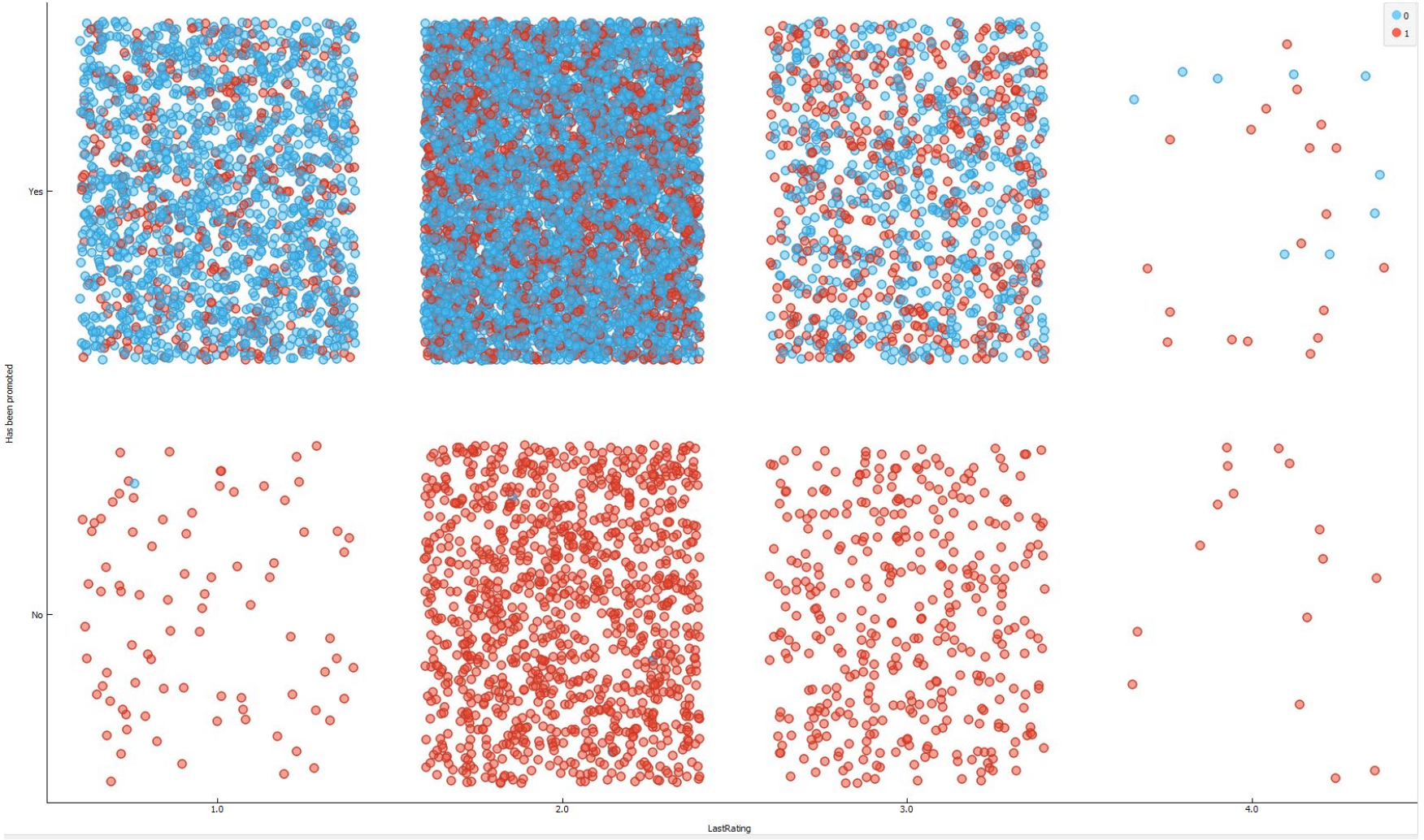
Employee Termination looks higher



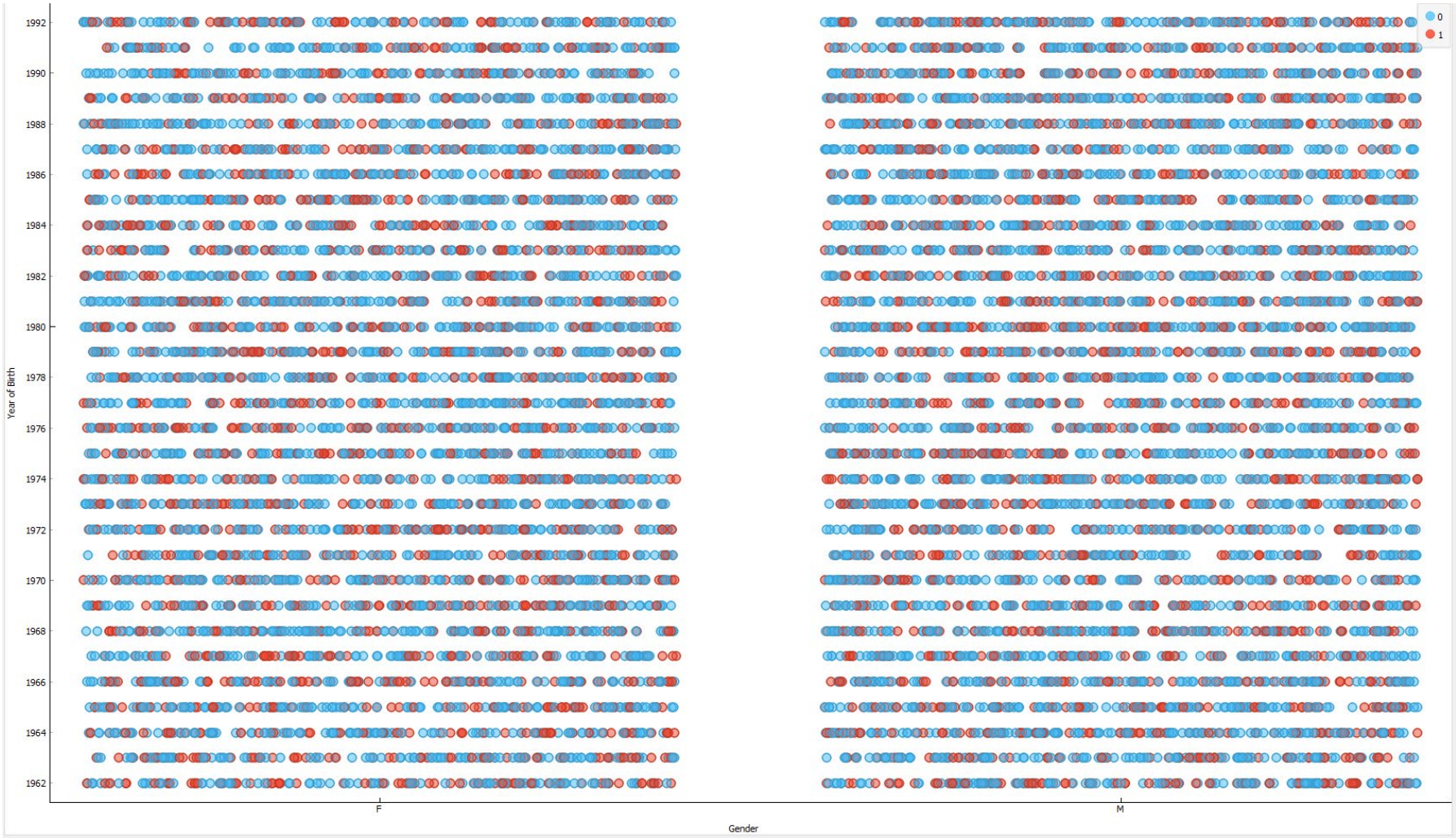
Salary could be Significant -> low salary can make person to exit



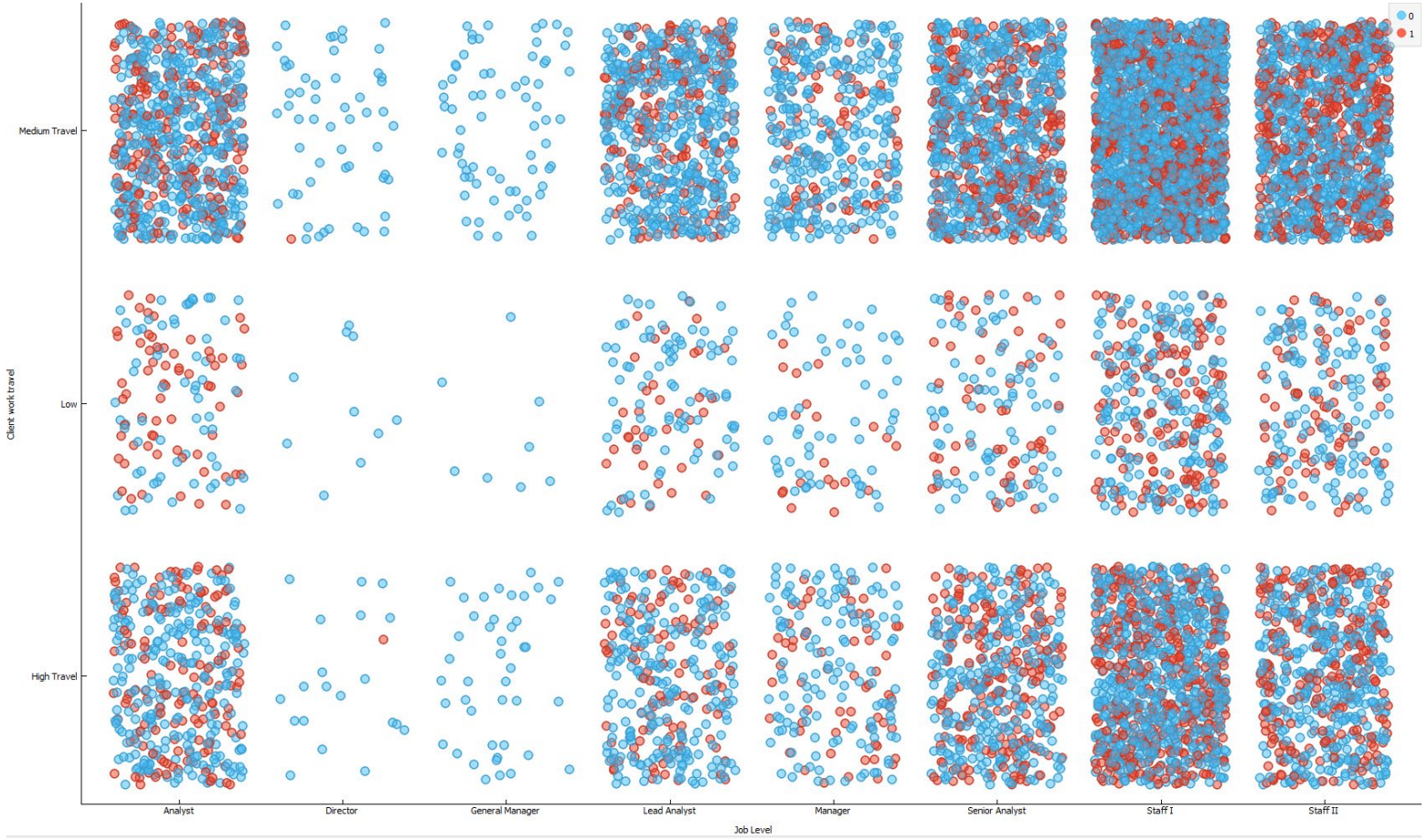
Rating - not much Significant (looks like terminations follow population i.e uniform distribution)



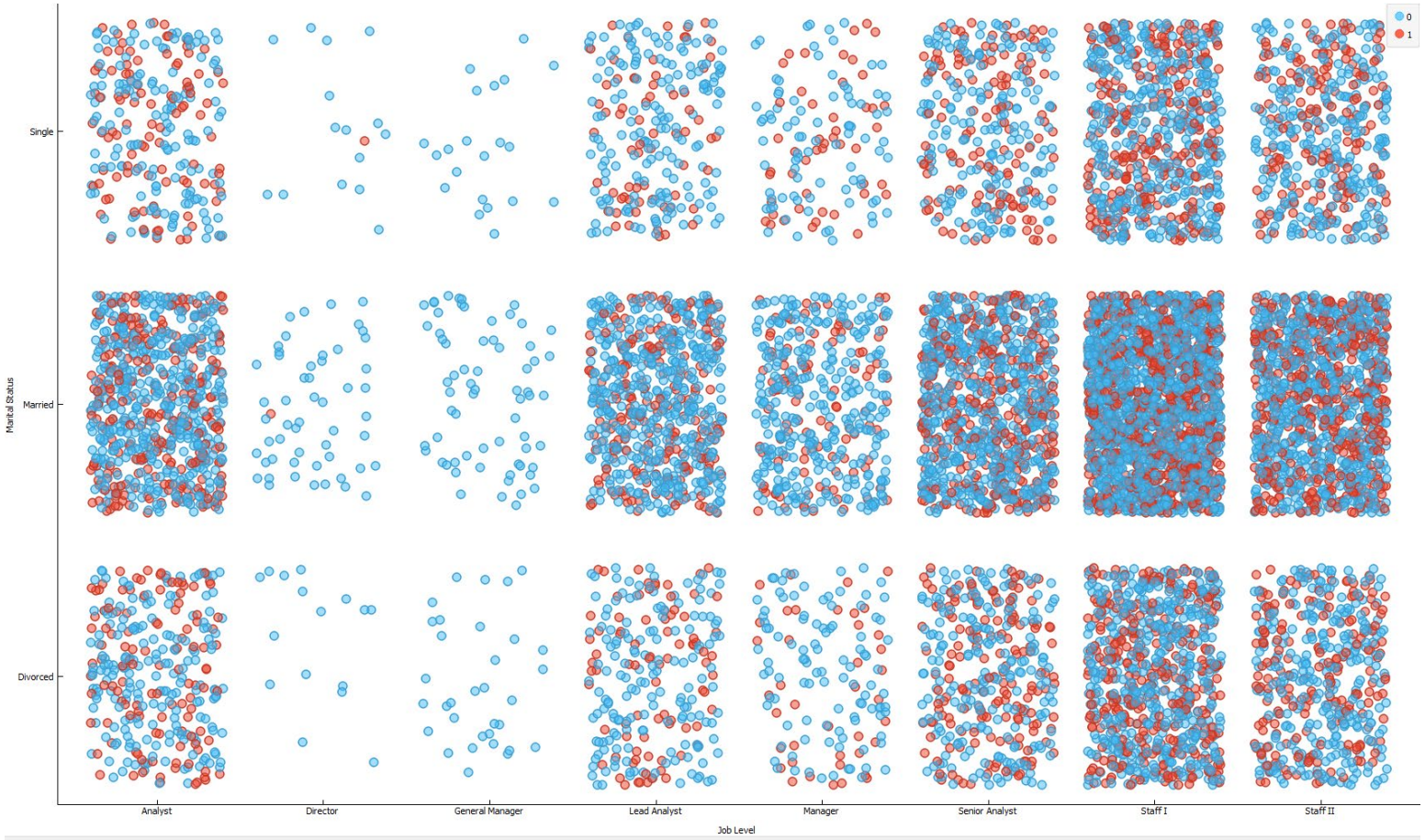
YOB and Gender not Significant (Uniform distribution)



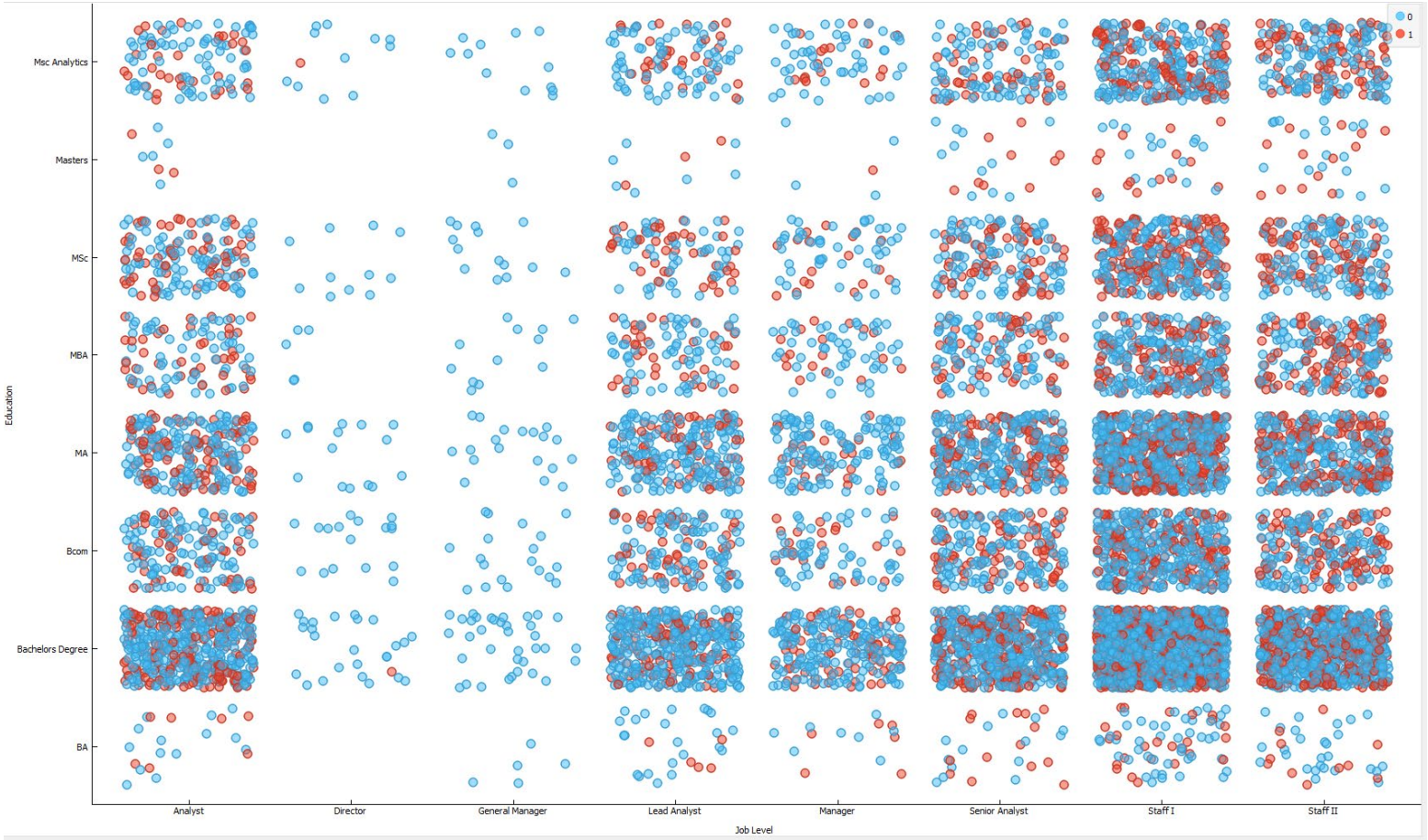
Travel not very significant (Uniform distribution)



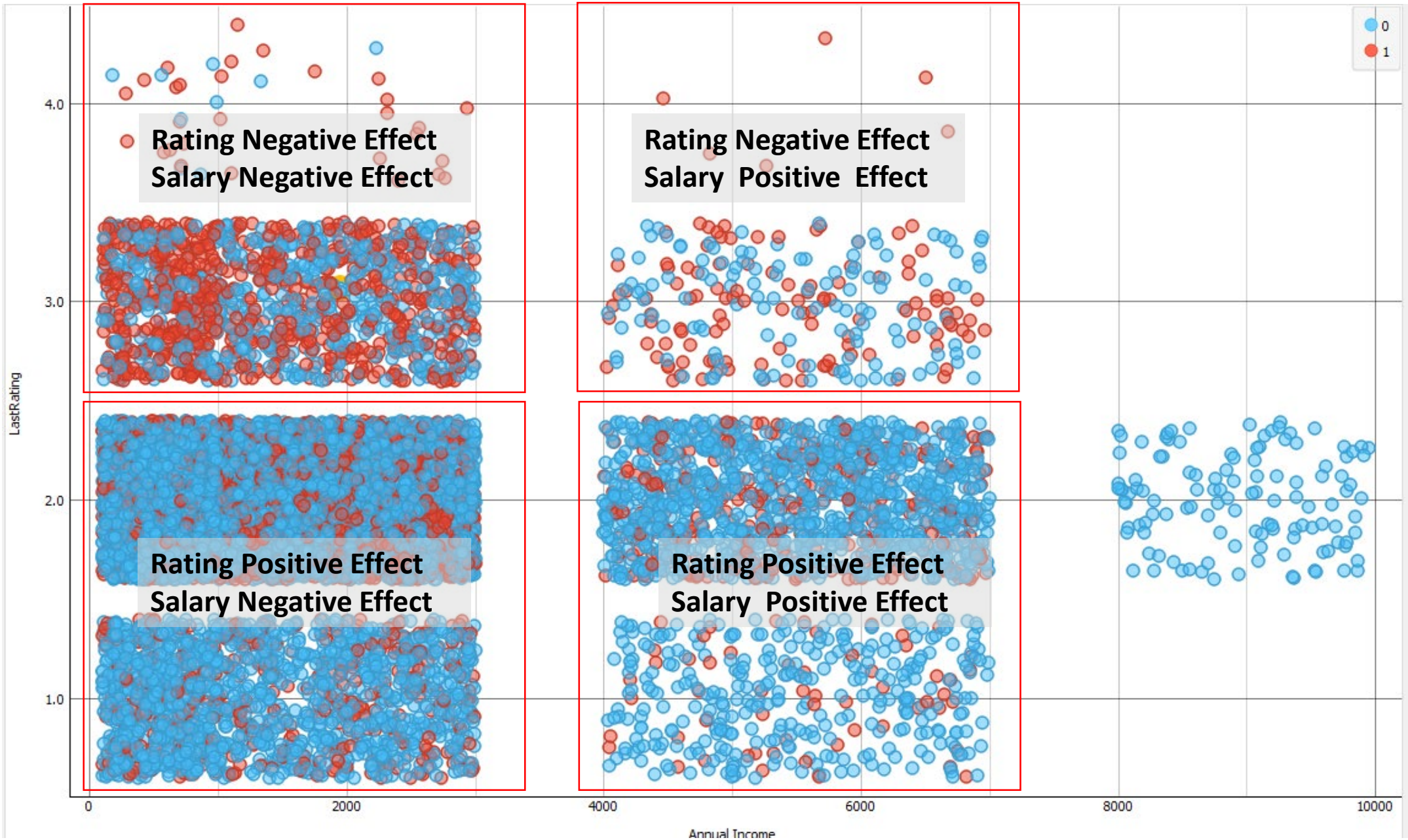
Martial Status not significant (Uniform distribution)



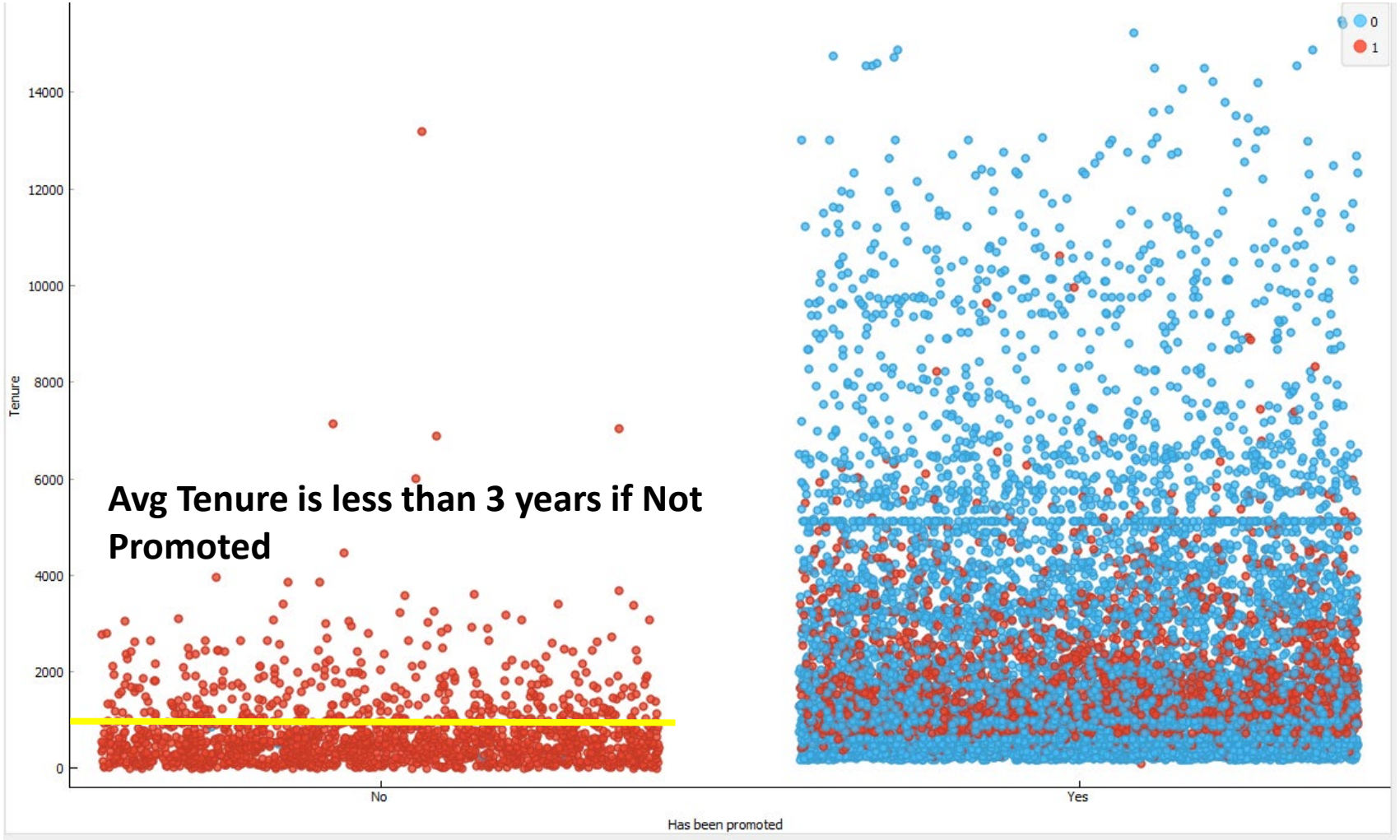
Education Significance could also be less (Uniform distribution)



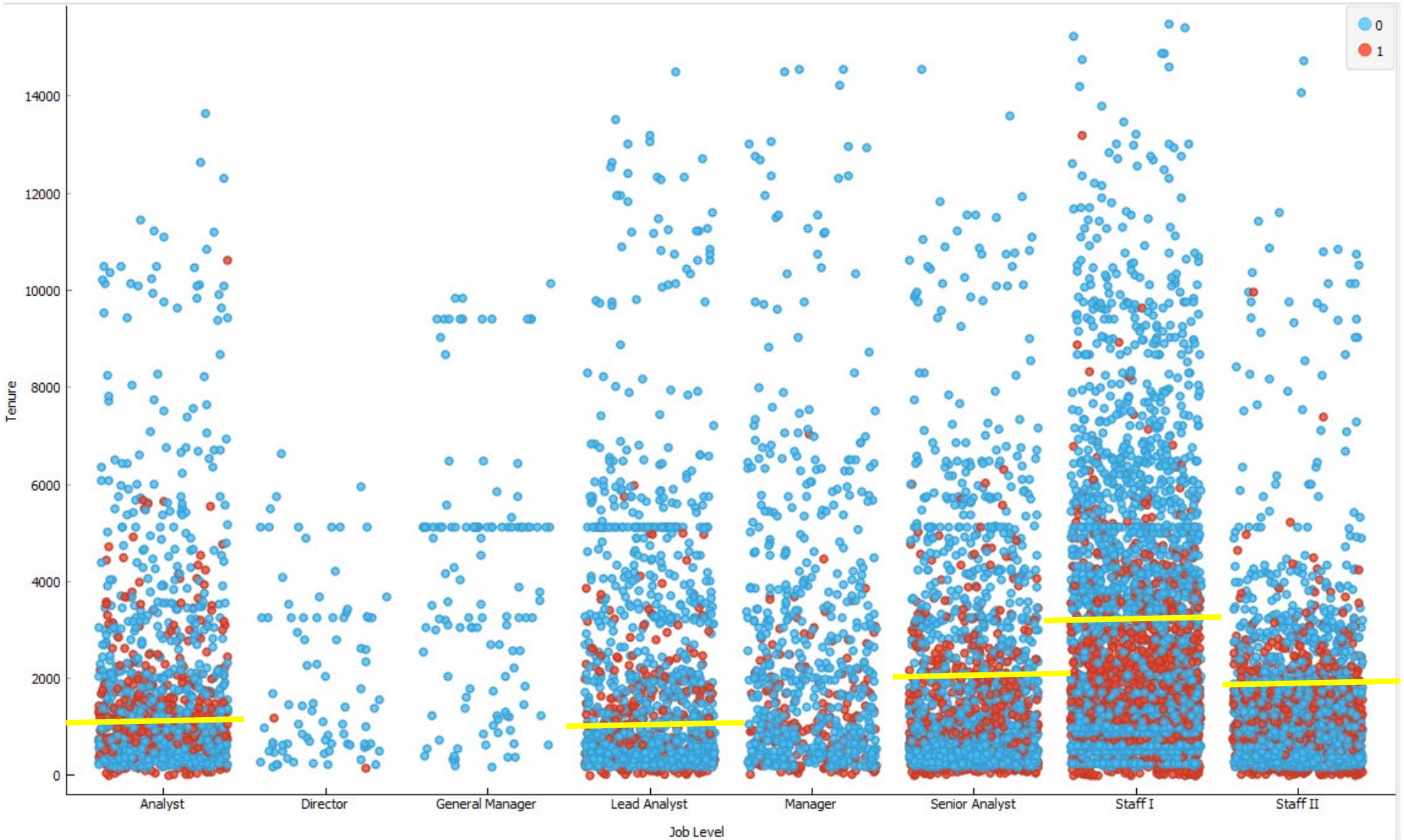
Combined effects of Rating and salary on exit of employee



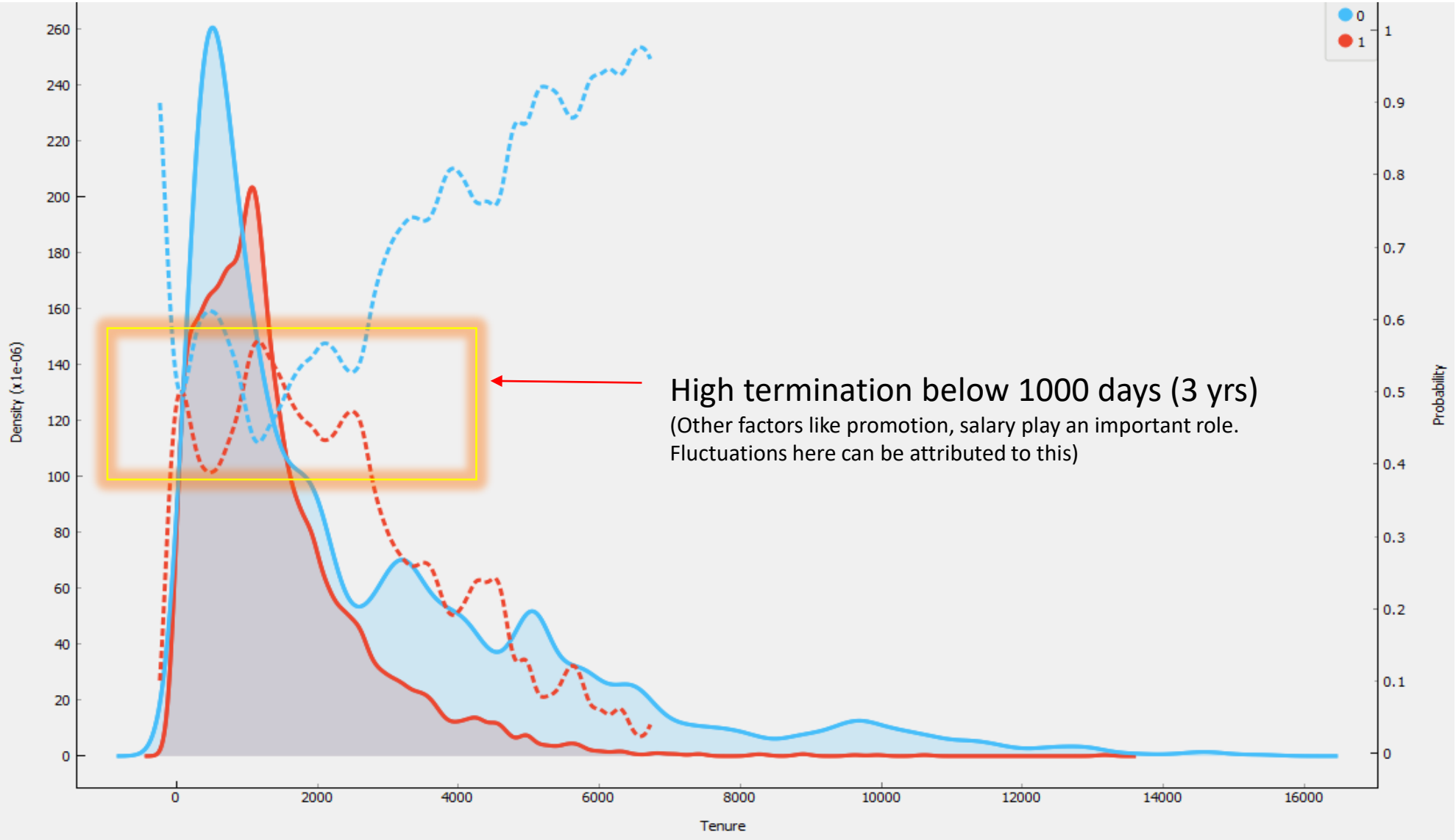
Effect of Promotion on Tenure and exit of employee



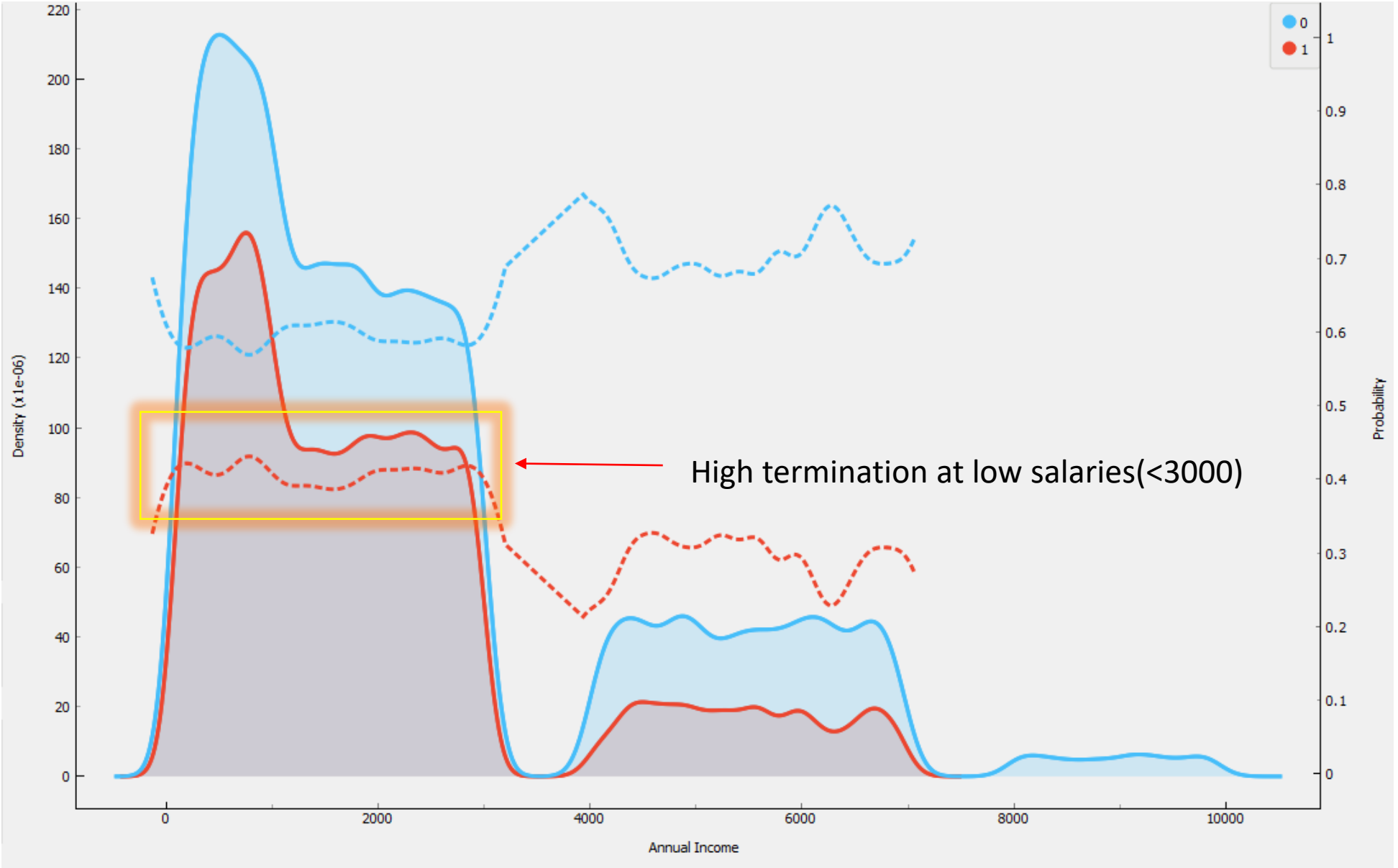
Tenure(days) Distribution across various job levels



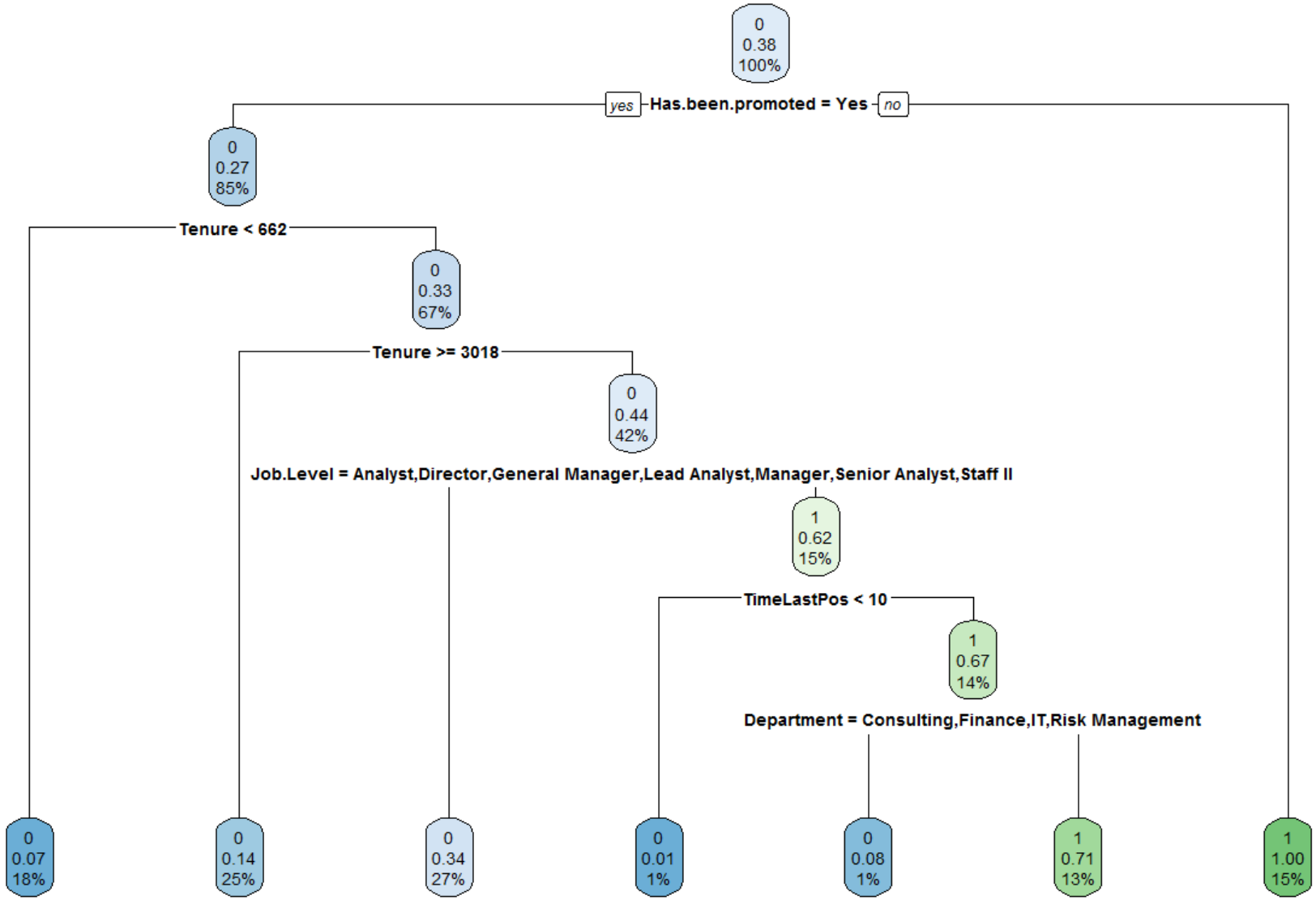
Tenure vs exit



Annual Income vs exit



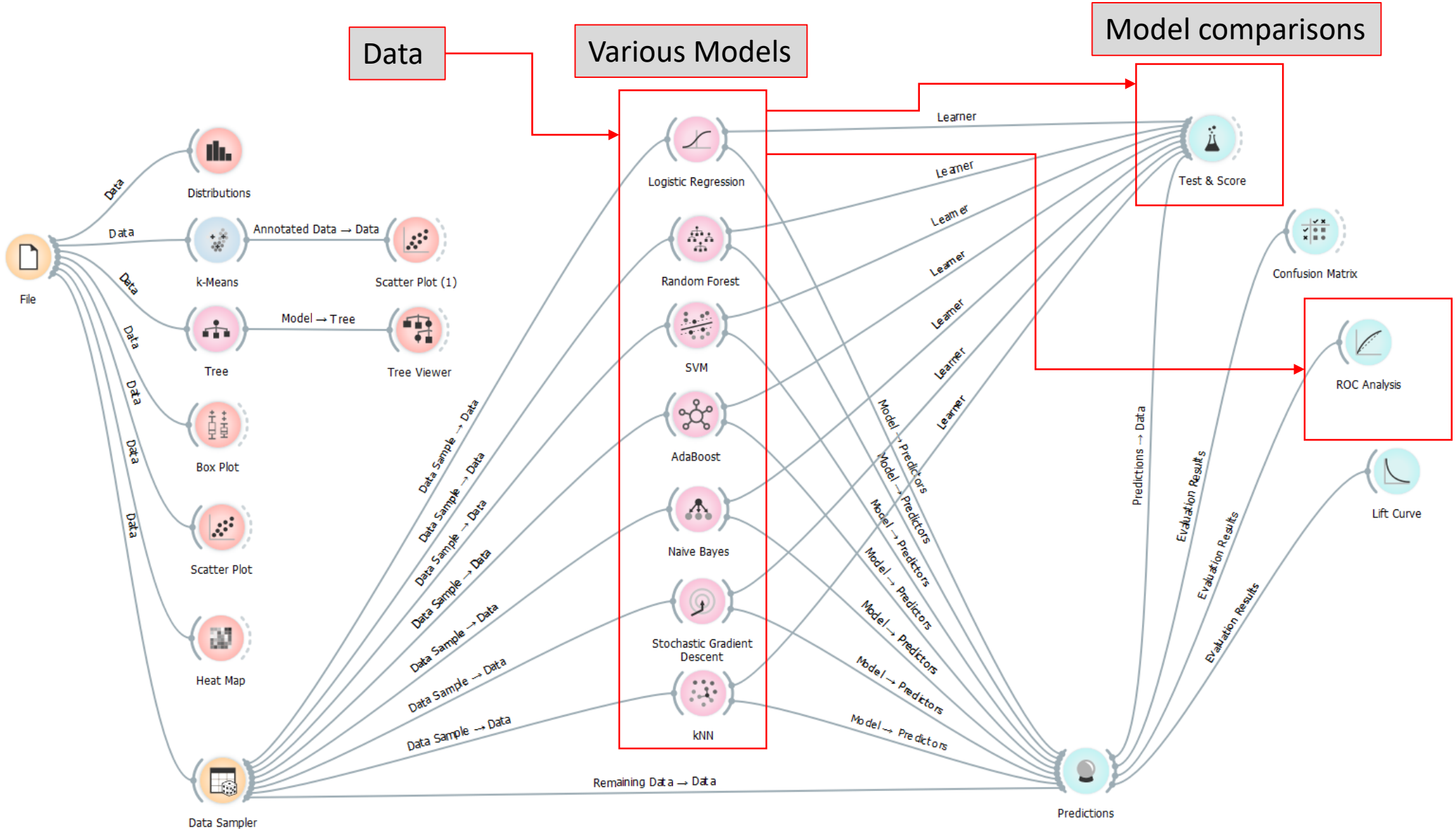
Basic Decision tree (part of exploratory analysis)



Insights from Exploratory data analysis

- Promotion, Tenure, Job levels, Department, Annual income are significant
- Demographic variables like gender, Year of birth, marital status are less significant. Same goes with travel time.
- Promotion turns out to be a key parameter in assessing the risk profile.
  - Promotion in High and mid level management positions could significantly reduce attrition in this segment.
  - Along with promotion, salary turns out to be an important driver at entry level jobs to retain them
- All these insights have to be validated using Machine learning algorithms.

Choosing a right classification model



Choosing a right classification model

Model comparisons

Method	AUC	CA	F1	Precision	Recall
Random Forest	0.890	0.822	0.740	0.816	0.676
Naive Bayes	0.843	0.791	0.703	0.754	0.658
Logistic Regression	0.822	0.782	0.628	0.873	0.491
AdaBoost	0.796	0.810	0.745	0.747	0.744
Tree	0.786	0.819	0.752	0.773	0.731
kNN	0.755	0.709	0.588	0.627	0.554
Stochastic Gradient Descent	0.699	0.767	0.578	0.900	0.426
SVM	0.514	0.564	0.352	0.397	0.316

Area Under the Curve (AUC)

Area under ROC curve is often used as a measure of quality of the classification models. A random classifier has an area under the curve of 0.5, while AUC for a perfect classifier is equal to 1. In practice, most of the classification models have an AUC between 0.5 and 1.

- .90-1 = excellent (A)
- .80-.90 = good (B)
- .70-.80 = fair (C)
- .60-.70 = poor (D)
- .50-.60 = fail (F)

Accuracy(CA): the proportion of the total number of predictions that were correct.

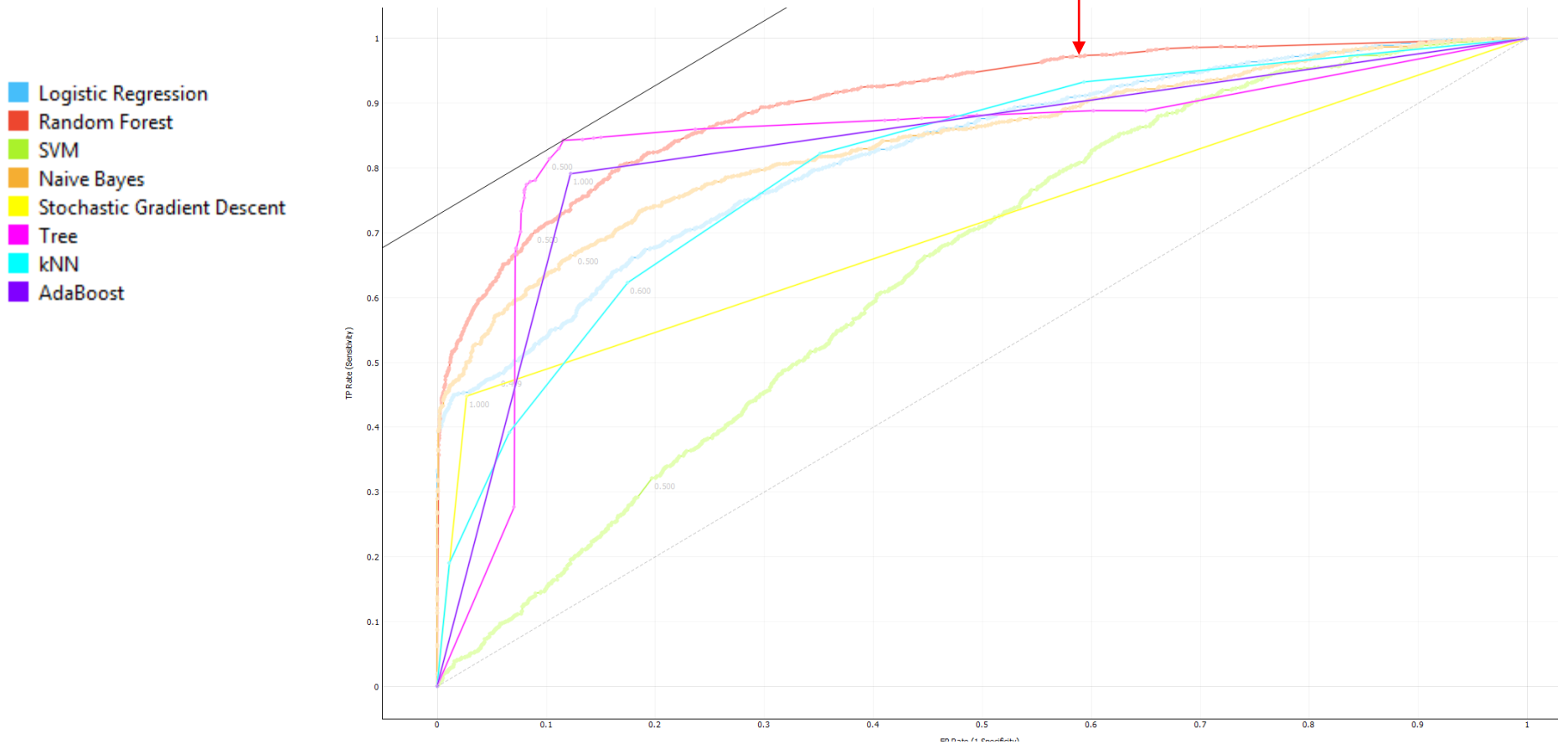
Precision : the proportion of positive cases that were correctly identified.

Sensitivity or Recall : the proportion of actual positive cases which are correctly identified.

Choosing a right classification model

ROC - AUC Analysis

Random forest is performing well in a balanced way



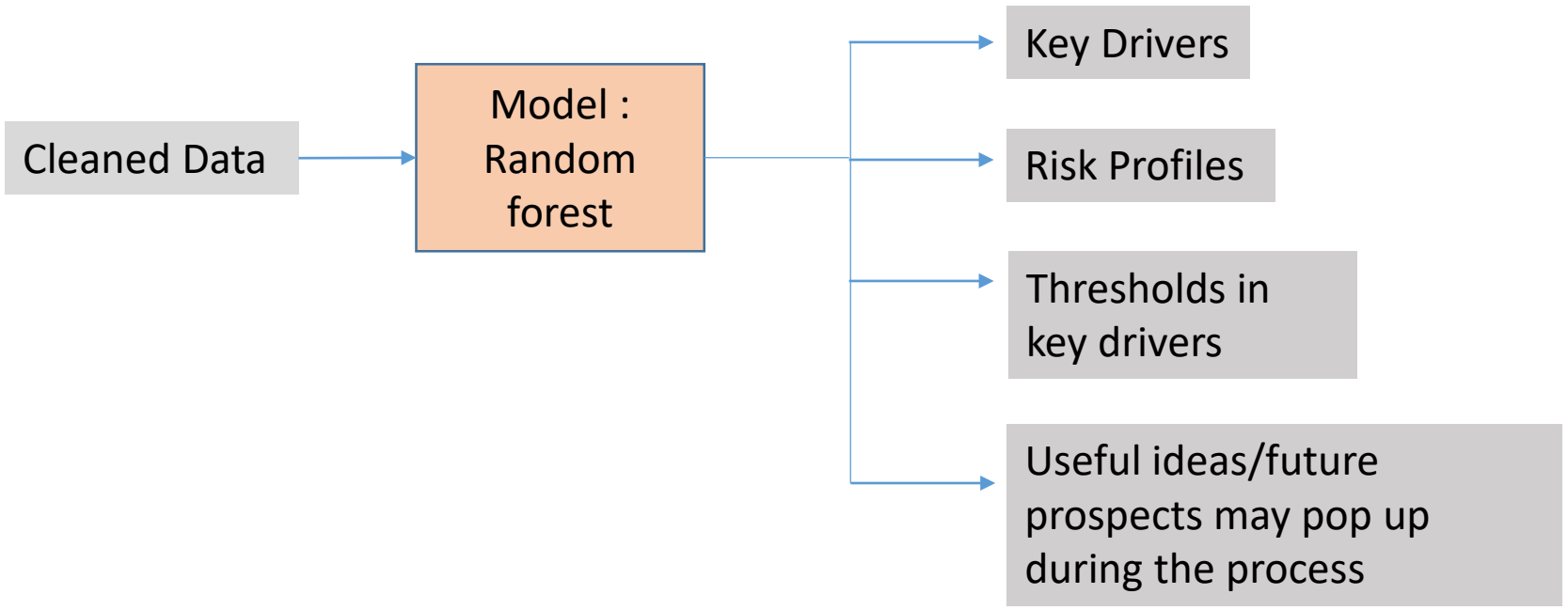
Choosing a right classification model

By removing less significant variables (from exploratory analysis) and running the model few improvements in KPIs is seen

Method	AUC	CA	F1	Precision	Recall
Random Forest	0.901	0.824	0.745	0.821	0.681
Naive Bayes	0.833	0.770	0.675	0.722	0.634
Logistic Regression	0.817	0.769	0.597	0.871	0.454
AdaBoost	0.813	0.825	0.767	0.771	0.762
kNN	0.760	0.711	0.596	0.631	0.564
Stochastic Gradient Descent	0.696	0.771	0.563	0.998	0.392
SVM	0.536	0.597	0.369	0.449	0.313

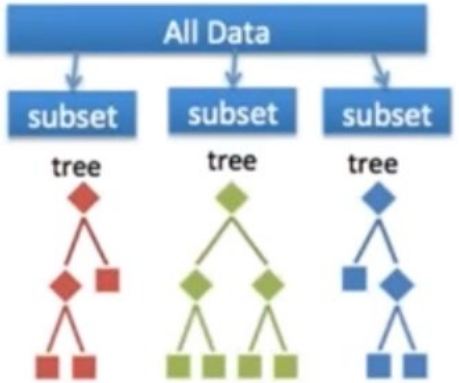
Model chosen: Random forest

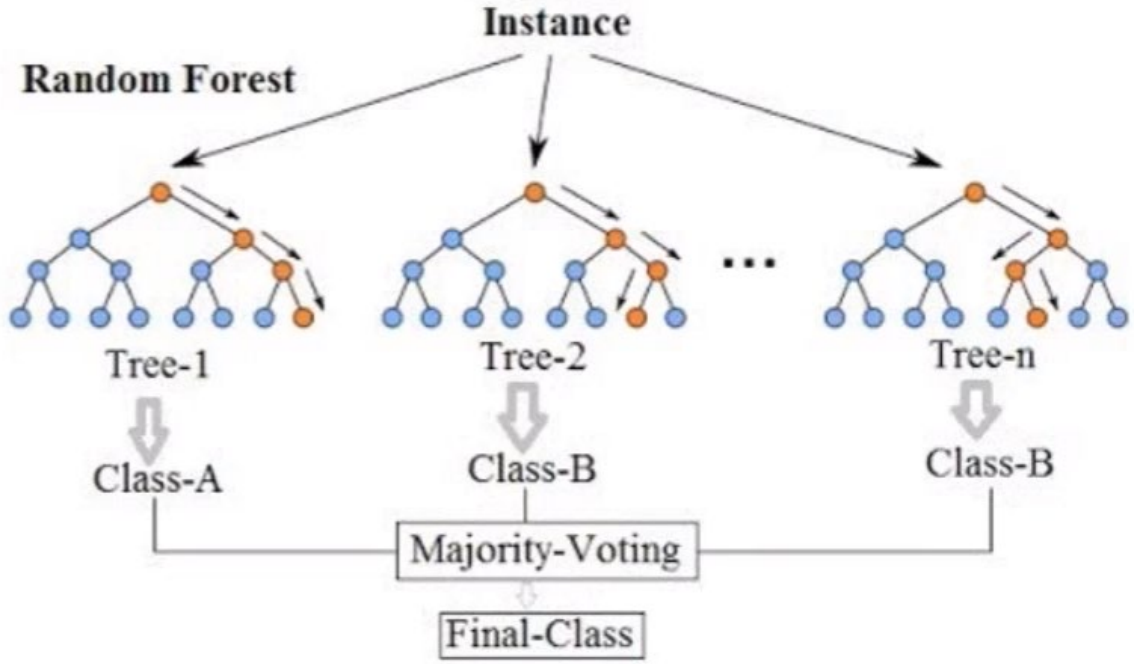
Objectives of model



Random forest (a brief overview)

- Random Forest is an ensemble classifier made using many decision tree models.
- Ensemble models combine the results from different models.





Random forest (a brief overview)

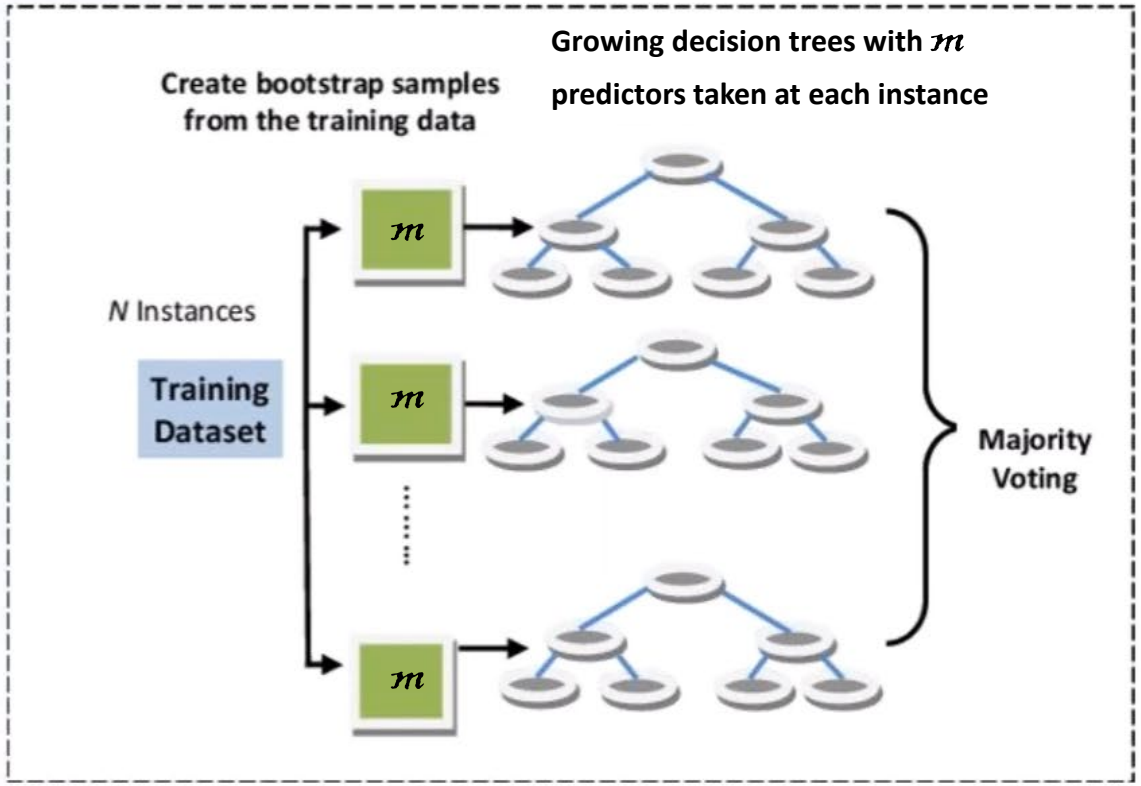
Randomly select m features from T; where $m \ll T$

For node d, calculate the best split point among the m feature

Split the node into two daughter nodes using the best split

Repeat first three steps until n number of nodes has been reached

Build your forest by repeating steps i–iv for D number of times



- ✓ T: number of features
- ✓ D: number of trees to be constructed
- ✓ V: Output: the class with the highest vote

- Data Acquisition
- Divide dataset
- Implement model
- Visualize
- Model Validation

```
#Loading the Data into R
hr_data<-read.csv("HR Sample V2.csv")
hr_data$Terminated<-as.factor(hr_data$Terminated)
```

```
> str(hr_data)
'data.frame':   10420 obs. of  15 variables:
 $ Rehire            : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
 $ Terminated        : Factor w/ 2 levels "0","1": 1 2 2 2 2 2 2 2 2 2 ...
 $ Employee_code     : int  77106 42830 67966 124106 12478 46454 16896 90530 11838 137958 ...
 $ Department        : Factor w/ 10 levels "","Audit","Consulting",..: 2 2 7 4 4 10 2 5 4 4 ...
 $ Job.Level         : Factor w/ 8 levels "Analyst","Director",..: 7 7 7 8 1 7 7 1 8 7 ...
 $ Tenure            : int  4119 1 2 2 3 5 5 5 5 5 ...
 $ TimeLastPos       : int  3840 1 2 2 3 5 5 5 5 5 ...
 $ Has.been.promoted : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 1 1 1 1 ...
 $ LastRating        : int  2 2 2 2 2 2 2 2 2 2 ...
 $ Client.work.travel: Factor w/ 3 levels "High Travel",..: 3 3 3 3 3 2 3 3 3 3 ...
 $ Education         : Factor w/ 8 levels "BA","Bachelors Degree",..: 4 2 3 3 3 2 4 2 4 2 ...
 $ Gender            : Factor w/ 2 levels "F","M": 1 2 2 2 2 1 2 2 2 2 ...
 $ Marital.Status    : Factor w/ 3 levels "Divorced","Married",..: 2 2 2 2 2 3 2 2 2 2 ...
 $ Annual.Income     : int  380 187 193 2653 1450 728 204 2893 1761 356 ...
 $ Year.of.Birth     : int  1987 1981 1990 1974 1977 1962 1992 1980 1971 1990 ...
```

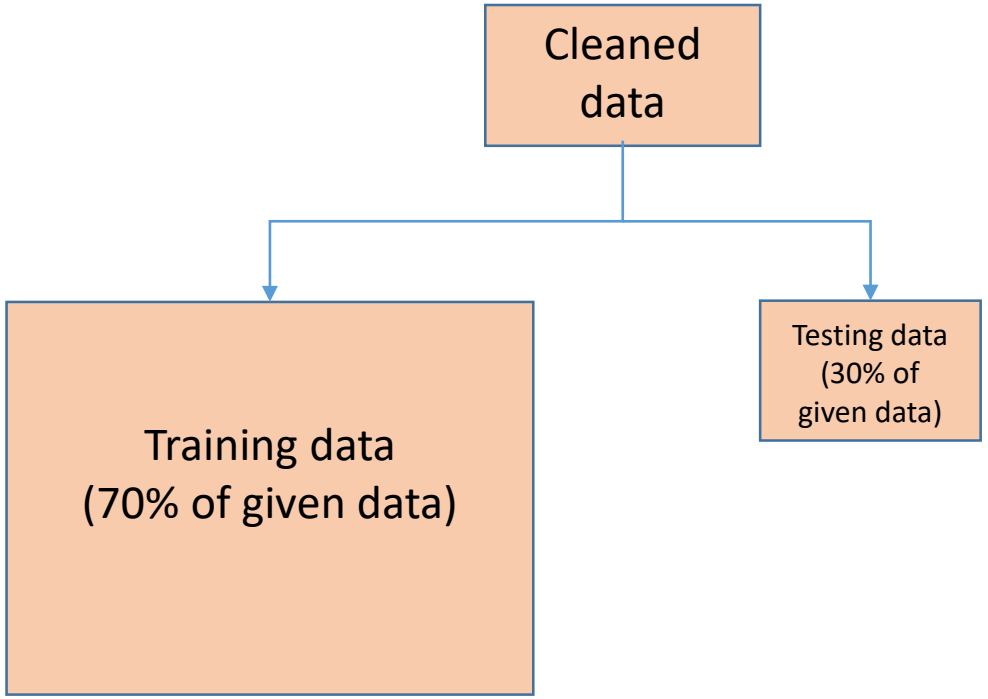
	Rehire	Terminated	Employee_code	Department	Job.Level	Tenure	TimeLastPos	Has.been.promoted	LastRating	Client.work.travel	Education	Gender	Marital.Status	Annual.Income	Year.of.Birth
1	FALSE	0	77106	Audit	Staff I	4119	3840	Yes	2	Medium Travel	MA	F	Married	380	1987
2	FALSE	1	42830	Audit	Staff I	1	1	No	2	Medium Travel	Bachelors Degree	M	Married	187	1981
3	FALSE	1	67966	Risk Management	Staff I	2	2	No	2	Medium Travel	Bcom	M	Married	193	1990
4	FALSE	1	124106	Finance	Staff II	2	2	No	2	Medium Travel	Bcom	M	Married	2653	1974
5	FALSE	1	12478	Finance	Analyst	3	3	No	2	Medium Travel	Bcom	M	Married	1450	1977
6	FALSE	1	46454	Tax	Staff I	5	5	No	2	Low	Bachelors Degree	F	Single	728	1962
7	FALSE	1	16896	Audit	Staff I	5	5	No	2	Medium Travel	MA	M	Married	204	1992
8	FALSE	1	90530	Financial Advisory	Analyst	5	5	No	2	Medium Travel	Bachelors Degree	M	Married	2893	1980
9	FALSE	1	11838	Finance	Staff II	5	5	No	2	Medium Travel	MA	M	Married	1761	1971

- Data Acquisition
- Divide dataset
- Implement model
- Visualize
- Model Validation

We will divide our entire dataset into two subsets as:

- Training dataset -> to train the model
- Testing dataset -> to validate and make predictions

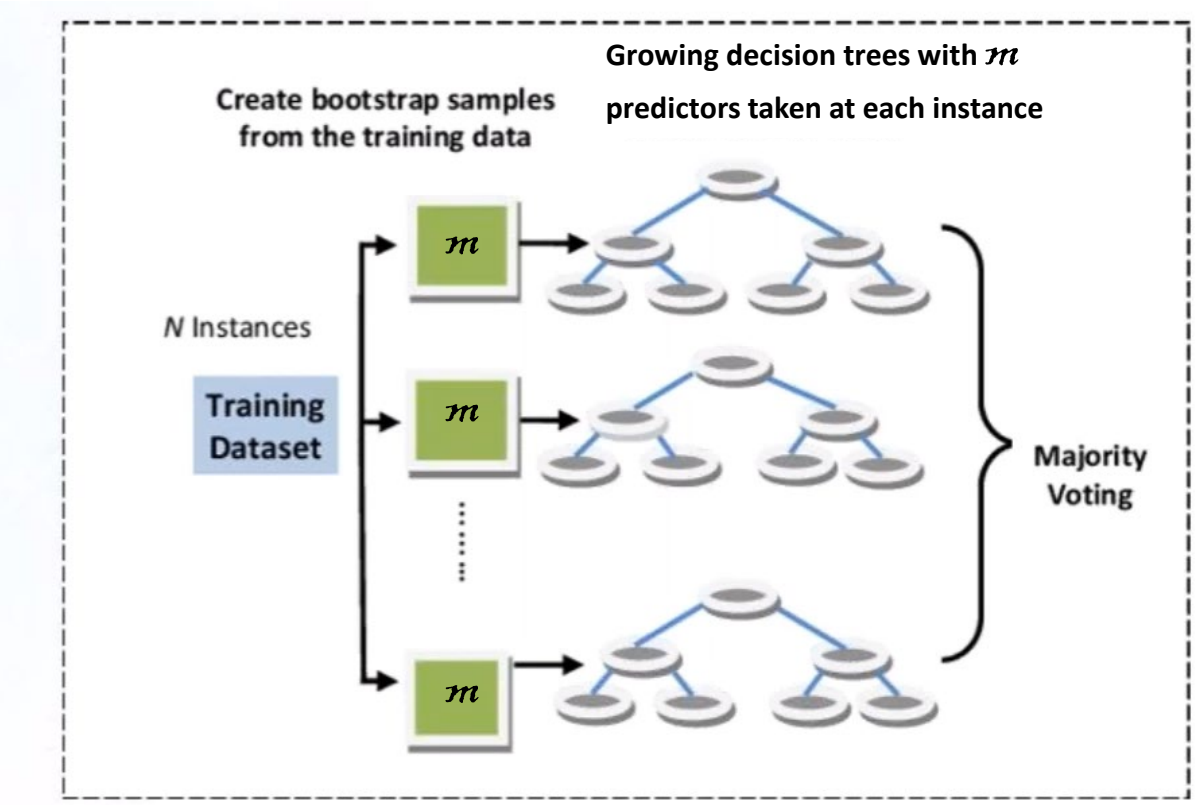
```
#library used for split funtion
library(caTools)
#Dividing the data into Training and Testing datasets
split <- sample.split(hr_data$Terminated, SplitRatio = 0.7)
training<-subset(hr_data,split=="TRUE")
testing<-subset(hr_data,split=="FALSE")
```



- Data Acquisition
- Divide dataset
- Implement model
- Visualize
- Model Validation

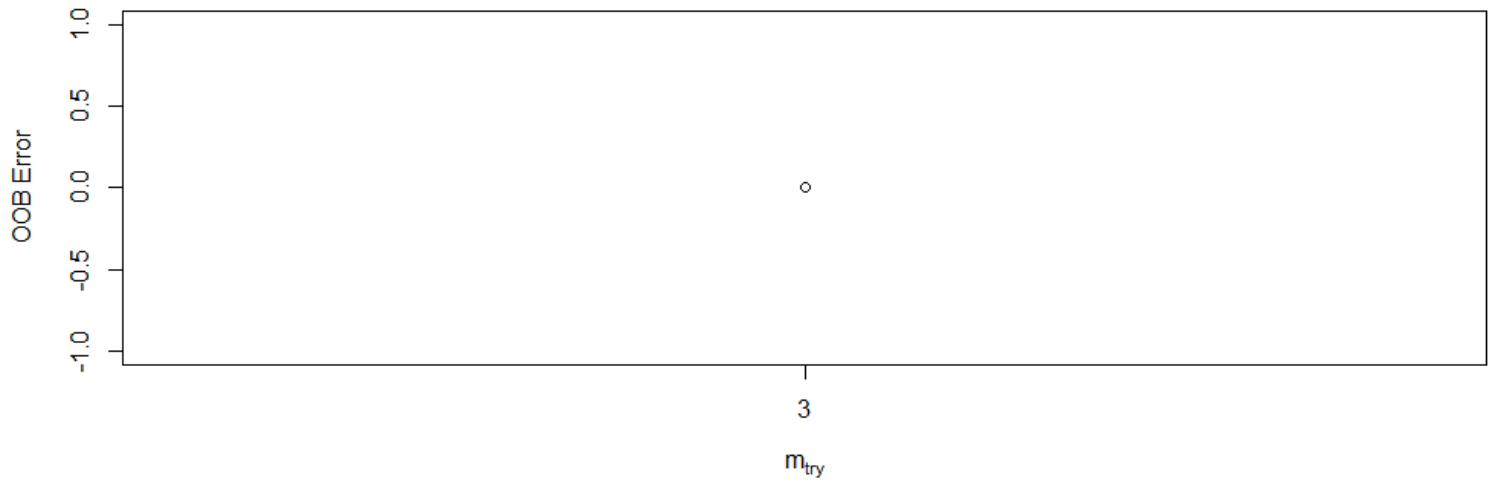
To find the value of m predictors to be taken at once in building the forest of decision trees

```
#optimised value of mtry
bestmtry <- tuneRF(training,training$Terminated,stepFactor = 1.2,improve = 0.05, trace = T, plot = T)
```



- Data Acquisition
- Divide dataset
- Implement model
- Visualize
- Model Validation

```
#optimised value of mtry
bestmtry <- tuneRF(training,training$Terminated,stepFactor = 1.2,improve = 0.05, trace = T, plot = T)
```



$m = 3$

i.e 3 predictors shall be suitable to be taken at once to build the forest of decision trees.

- Data Acquisition
- Divide dataset
- Implement model
- Visualize
- Model Validation

> Growing random forest

```
#Random Forest Library
library(randomForest)
### Model 1
Terminated_forest<-randomForest(Terminated~.-Employee_code,data = training)
Terminated_forest
```

> Results from random forest

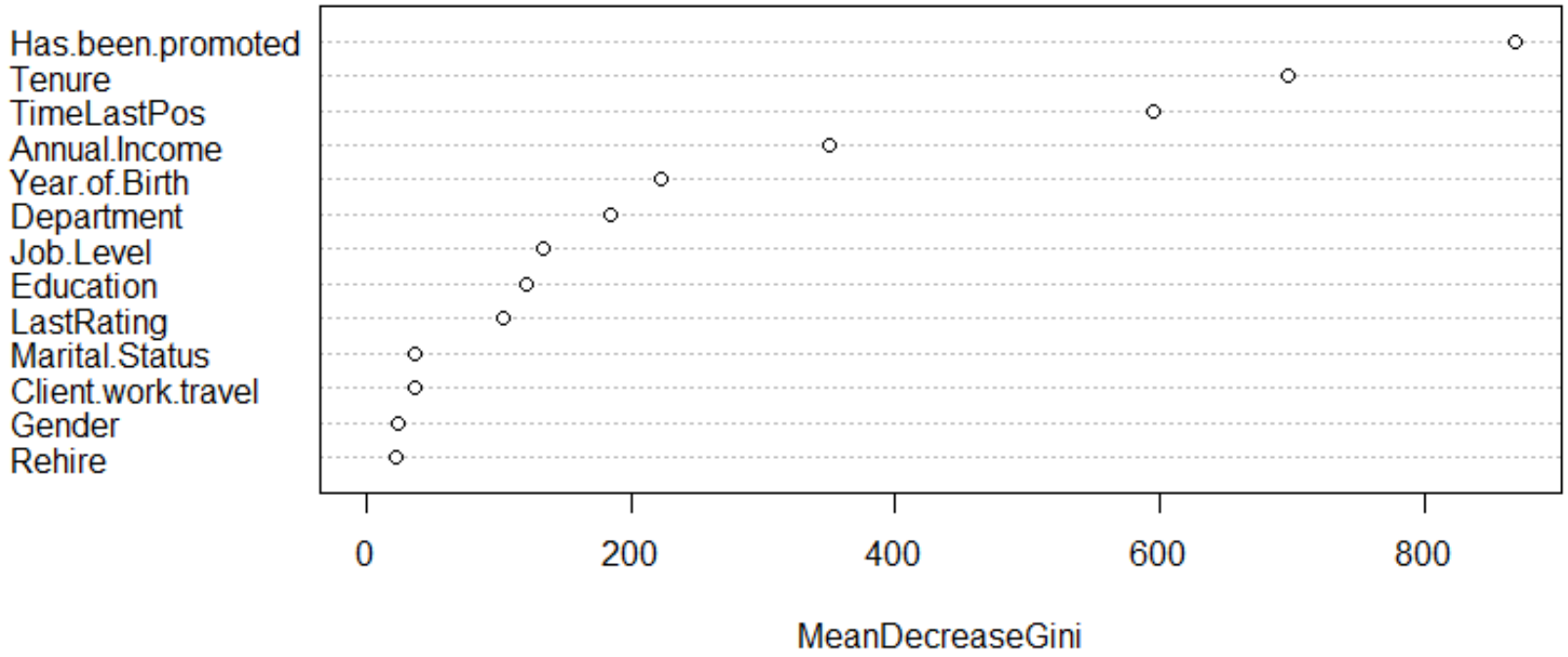
```
> Terminated_forest

Call:
 randomForest(formula = Terminated ~ . - Employee_code, data = training)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 15.5%
Confusion matrix:
     0    1 class.error
0 4329  360  0.07677543
1  818 2091  0.28119629
>
```

- Data Acquisition
- Divide dataset
- Implement model
- **Visualize**
- Model Validation

> Finding important predictors or key drivers of attrition from the model

```
#Gives Gini Index (Priority of variables)
importance(Terminated_forest)
varImpPlot(Terminated_forest)
```

> Plotting them



As per the model, *Has.been.promoted* is the most important predictor followed by Tenure TimeLastPos and so on till Rehire

- Data Acquisition
- Divide dataset
- Implement model
- Visualize
- **Model Validation**

Now, We can use our model to predict the outcome of our testing dataset.

```
#Prediction
Predrf <- predict(Terminated_forest, newdata=testing, type = "class")
Predrf

#validation
library(caret)
confusionMatrix(table(Predrf,testing$Terminated))
```

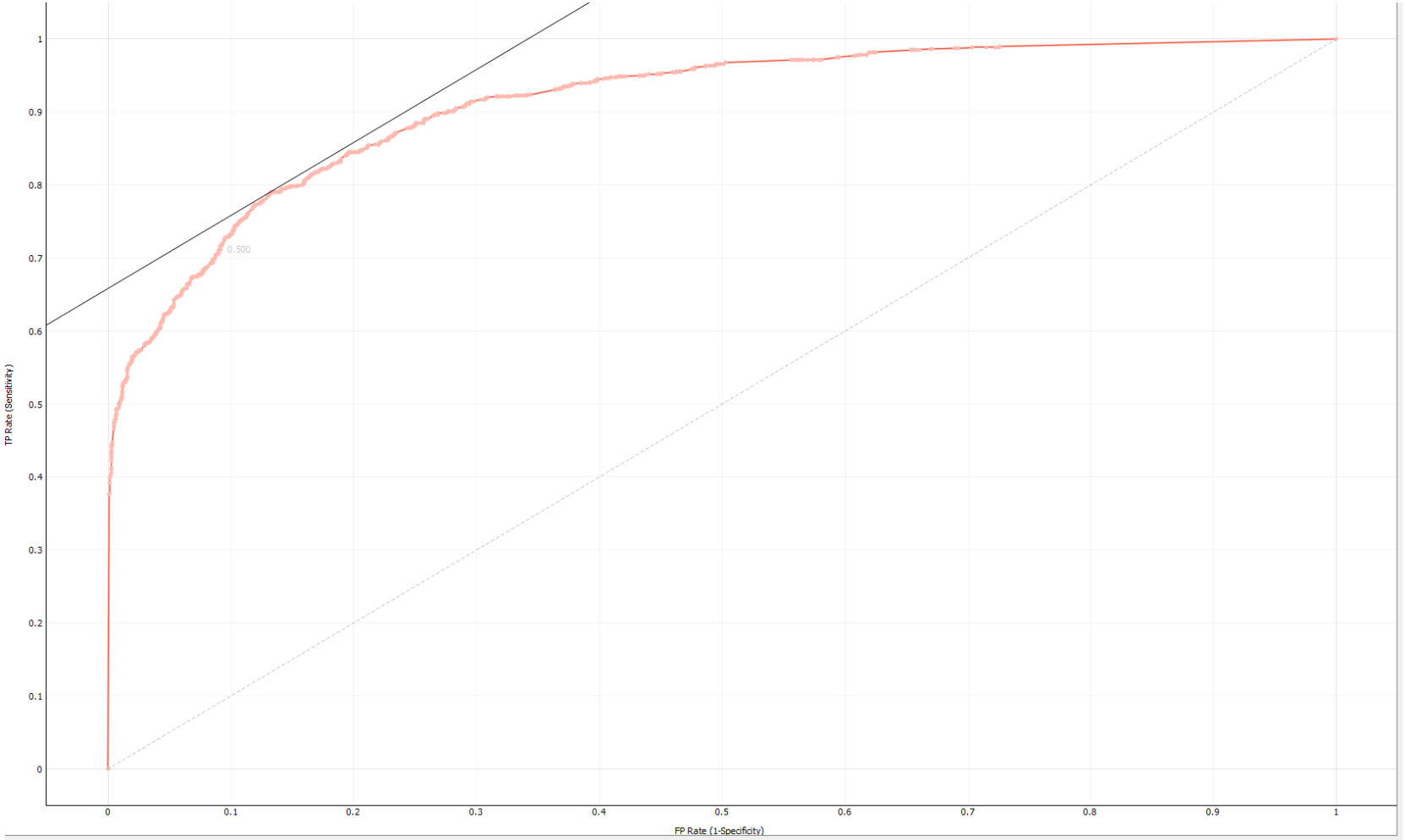
```
3054 3056 3060 3068 3069 3076 3078 3080 3088
   0    0    0    0    0    0    0    0    0
3089 3090 3097 3098 3101 3107 3111 3113 3116
   0    0    0    0    0    0    0    0    0
3117 3119 3121 3133 3135 3136 3137 3138 3139
   0    0    0    0    0    0    0    0    0
3140 3141 3142 3144 3147 3158 3159 3169 3172
   0    0    0    0    0    0    0    0    0
3173 3174 3179 3183 3187 3190 3195 3196 3198
   0    0    0    0    0    0    0    0    0
3199 3202 3204 3205 3209 3215 3218 3221 3222
   0    0    0    0    0    0    0    0    0
3230 3233 3234 3237 3243 3245 3247 3251 3255
   0    0    0    0    0    0    0    0    0
3258 3260 3263 3265 3270 3271 3276 3292 3295
   0    0    0    0    0    0    0    0    0
3300 3301 3304 3305 3320 3321 3322 3324 3325
   0    0    0    0    0    0    0    0    0
3335
   0
 [ reached getOption("max.print") -- omitted 2256 entries ]
Levels: 0 1
> |
```

```
Predrf    0    1
     0 1878  368
     1  131  879

               Accuracy : 0.8467
                 95% CI : (0.8339, 0.859)
    No Information Rate : 0.617
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6636
 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.9348
            Specificity : 0.7049
         Pos Pred Value : 0.8362
         Neg Pred Value : 0.8703
             Prevalence : 0.6170
         Detection Rate : 0.5768
   Detection Prevalence : 0.6898
      Balanced Accuracy : 0.8198

       'Positive' Class : 0
```

Accuracy of the model : 84.67%

- Data Acquisition
- Divide dataset
- Implement model
- Visualize
- **Model Validation**

ROC Curve



- Data Acquisition
- Divide dataset
- Implement model
- Visualize
- **Model Validation**

Confusion Matrix

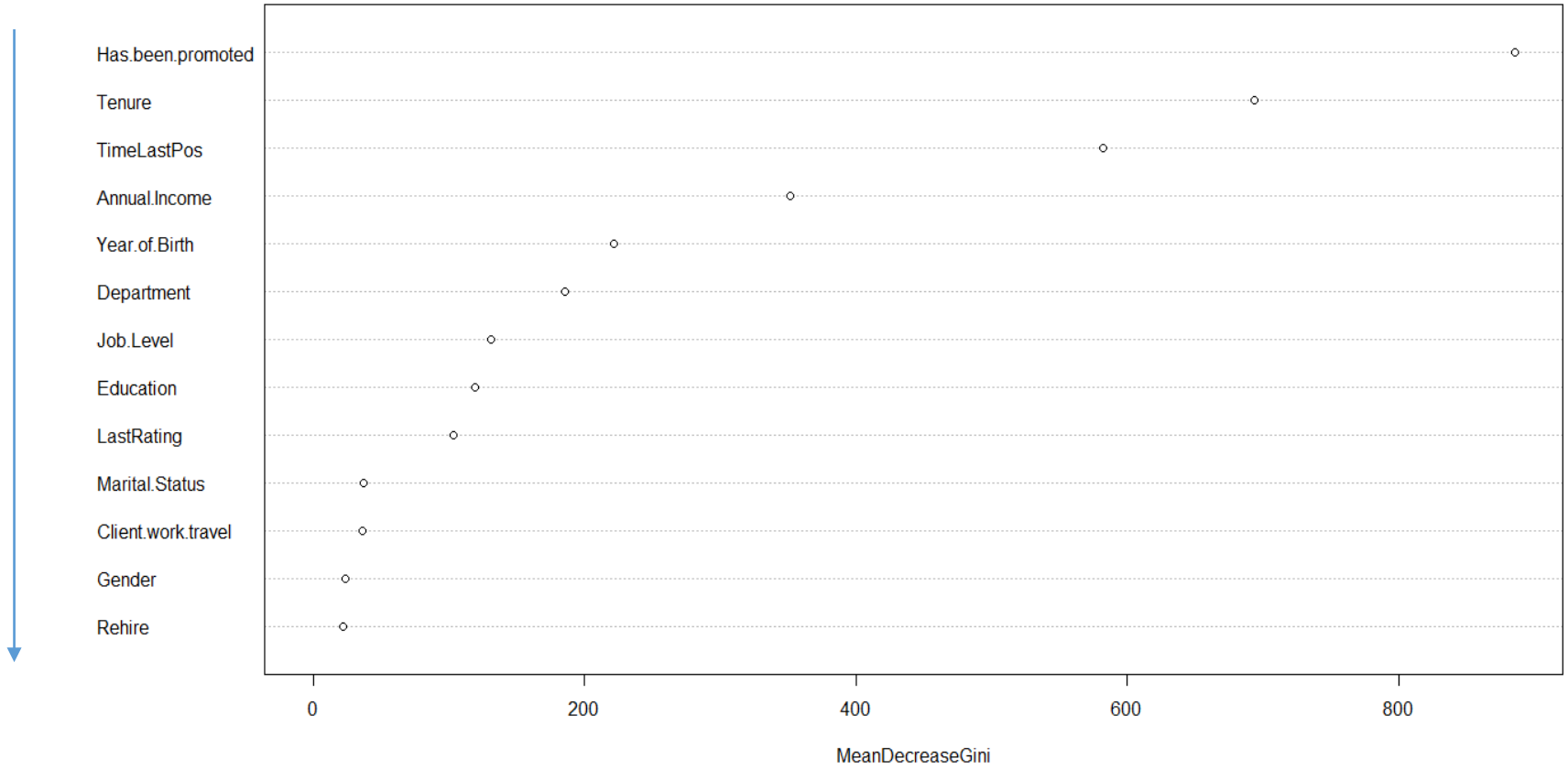
Random Forest		Predicted		
		0	1	
Actual	0	1878	368	0.836153161
	1	131	879	0.87029703
		0.93479343	0.70489174	0.846744472

- **Accuracy** : the proportion of the total number of predictions that were correct.
- **Positive Predictive Value or Precision** : the proportion of positive cases that were correctly identified.
- **Negative Predictive Value** : the proportion of negative cases that were correctly identified.
- **Sensitivity or Recall** : the proportion of actual positive cases which are correctly identified.
- **Specificity** : the proportion of actual negative cases which are correctly identified.

Confusion Matrix		Target			
		Positive	Negative		
Model	Positive	a	b	*Positive Predictive Value*	a/(a+b)
	Negative	c	d	*Negative Predictive Value*	d/(c+d)
		Sensitivity	*Specificity*	**Accuracy** = (a+d)/(a+b+c+d)	
		a/(a+c)	d/(b+d)		

1. Key Drivers for Termination or attrition

Decreasing order of importance



Has.been.promoted, Tenure, Time.LastPos, Annual income, Year.of.birth, Department, Job.Level turns out to be key drivers for termination or attrition

2. Risk Profiling

2. 1 Risk Profiling segmented approach

Probability of termination of each employee can be obtained from the model.
These can be divided into High, Medium and Low risk segments

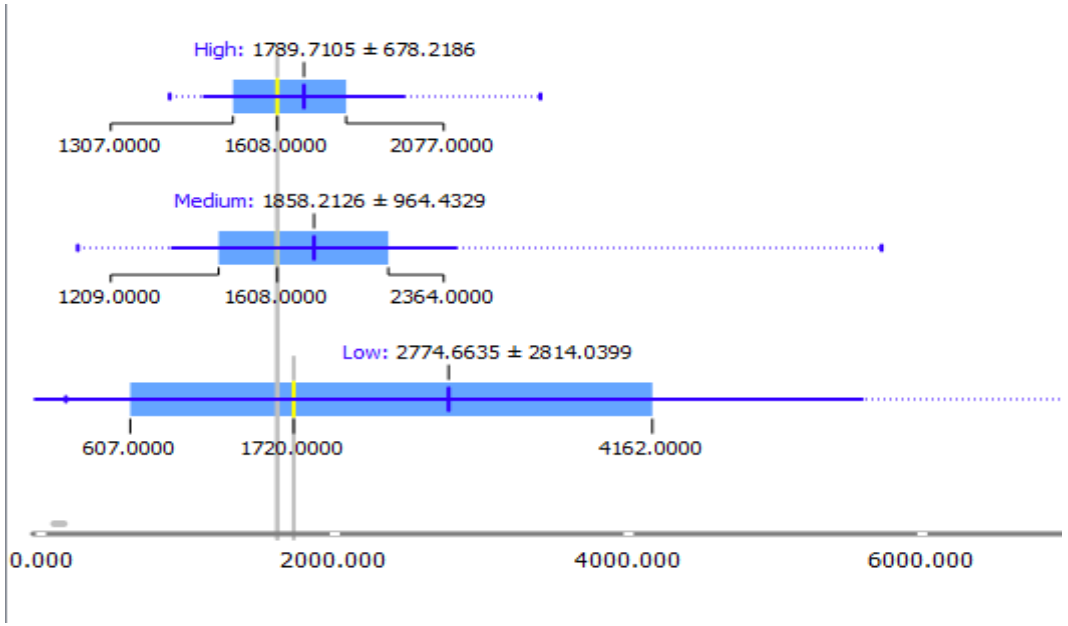
```
TestProbility <-as.data.frame(predict(Terminated_forest, newdata=hr_data, type = "prob"))
colnames(TestProbility)<-c("P0","P1")
NewSolution <-cbind(TestProbility,hr_data)
NewSolution<-NewSolution %>% mutate(Risk_class = ifelse(P1<= 0.33, "Low",
                                                        ifelse(P1>0.33 & P1<=0.66, "Medium", "High"
                                                        )))
```

```
> table(NewSolution$Risk_class,NewSolution$Terminated)

            0    1
  High     38 3268
  Low    5926  162
  Medium  475  551
```

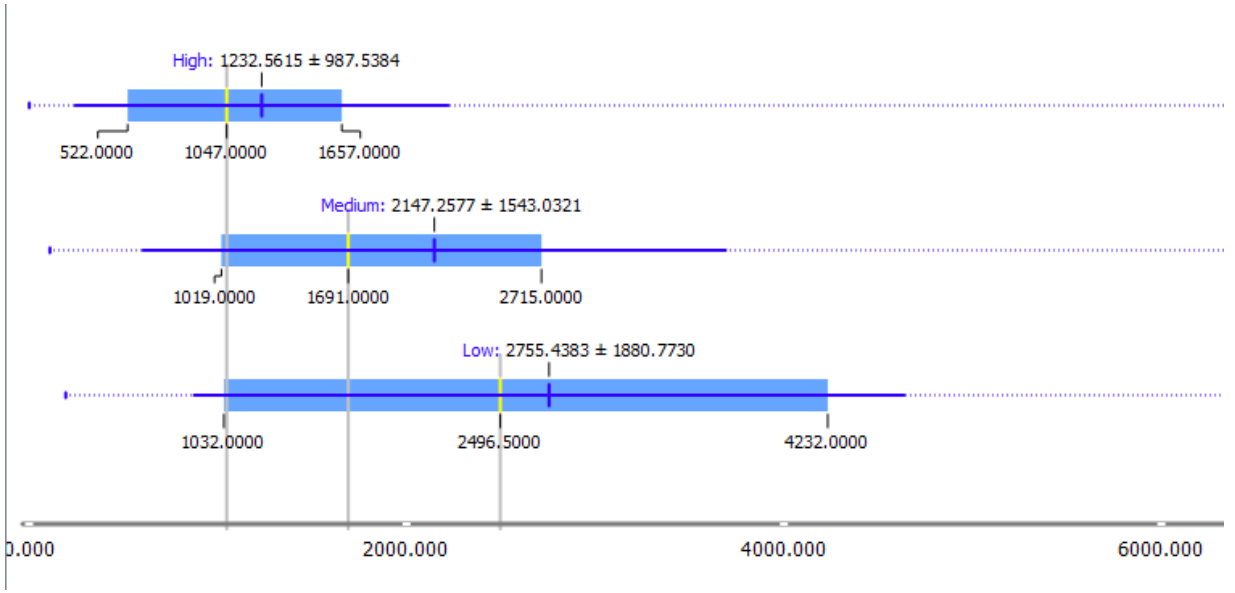
High → Error Rate : 1.14 %

Low → Error Rate : 2.45 %

2.2 Tenure thresholds for High, medium and low risk divisions

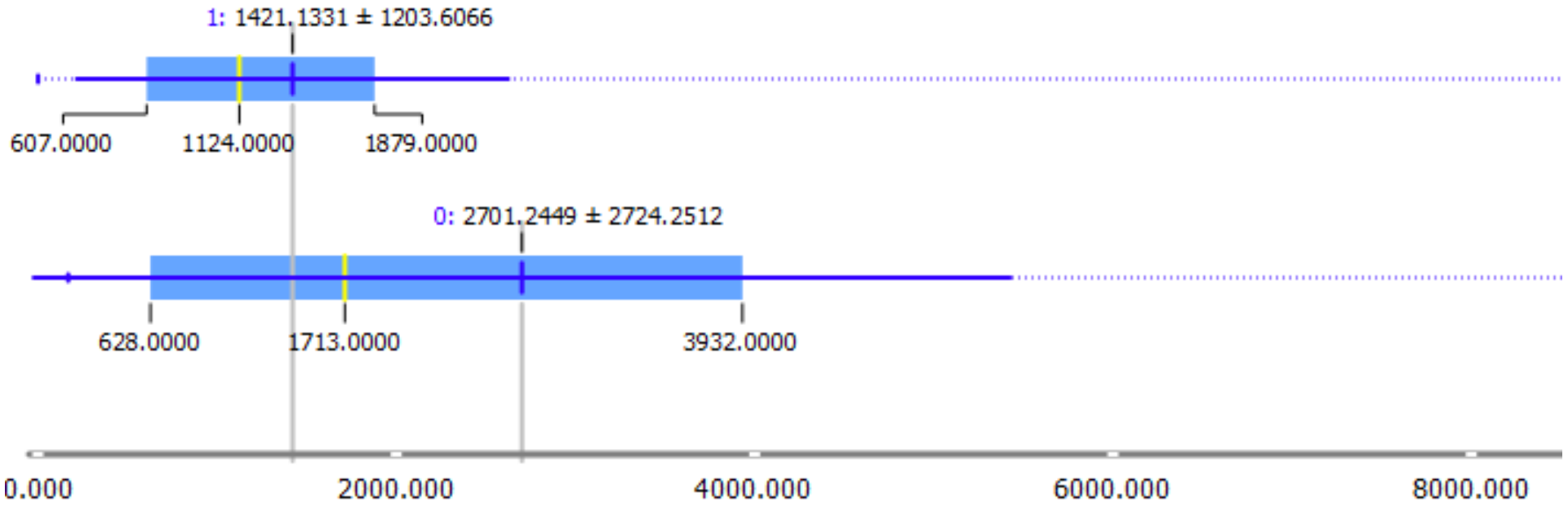


Terminated



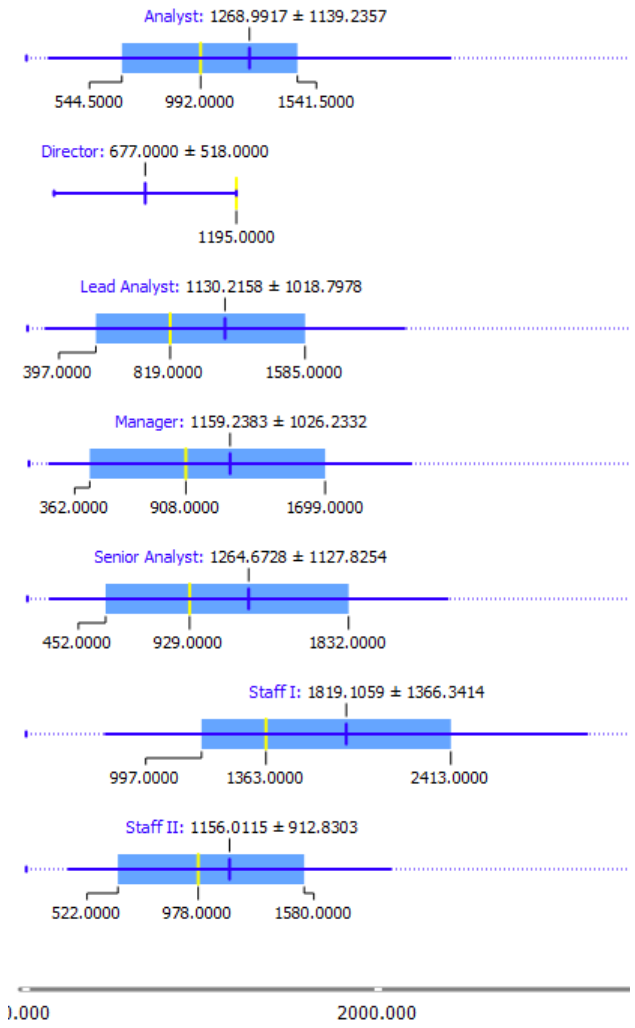
3. Thresholds - Tenure

3.1 Terminated Thresholds

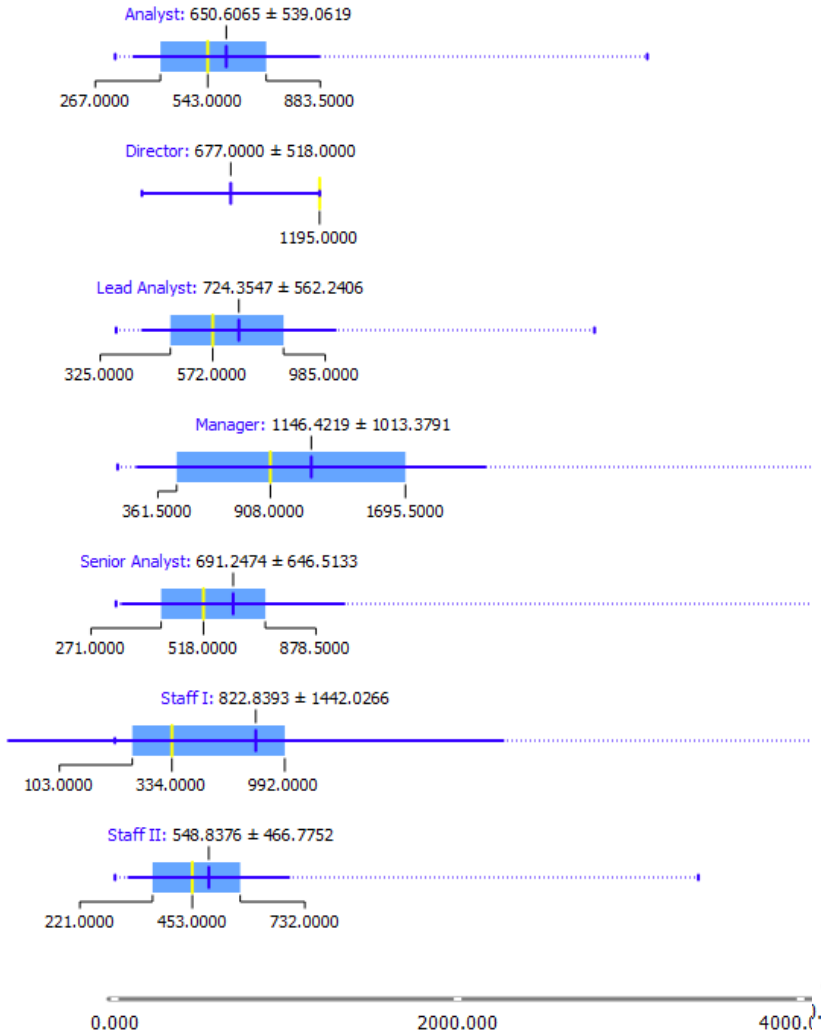


3.2 Job Profile Thresholds

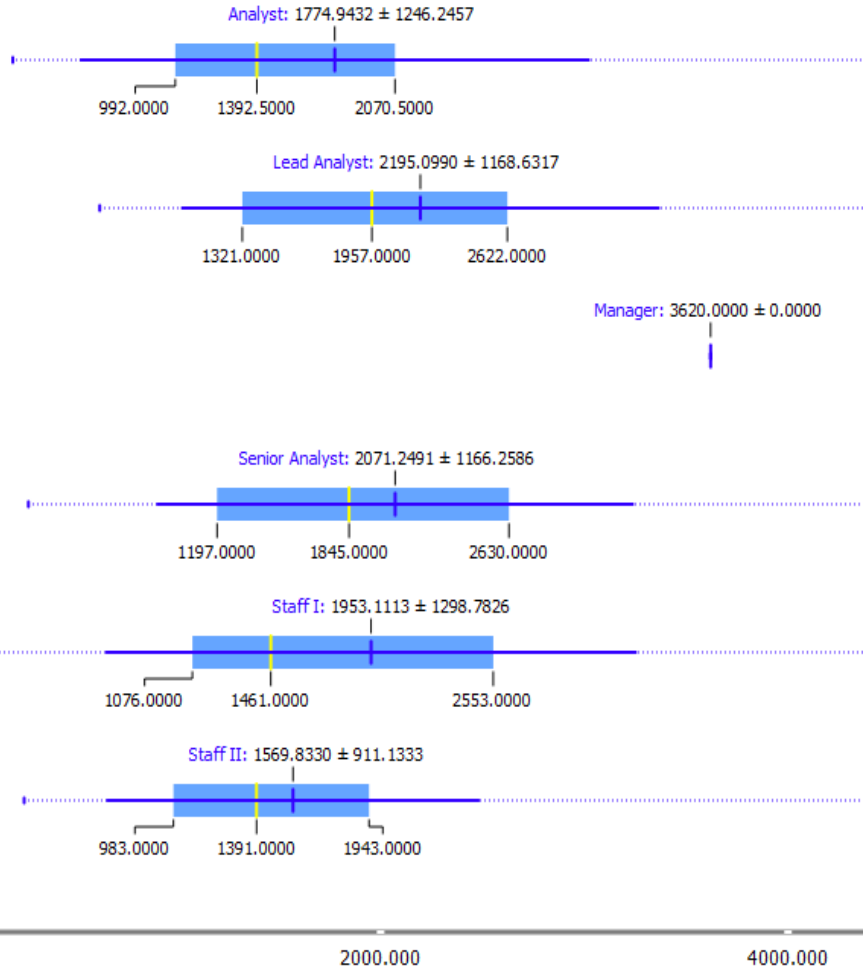
Terminated



Terminated – Not Promoted

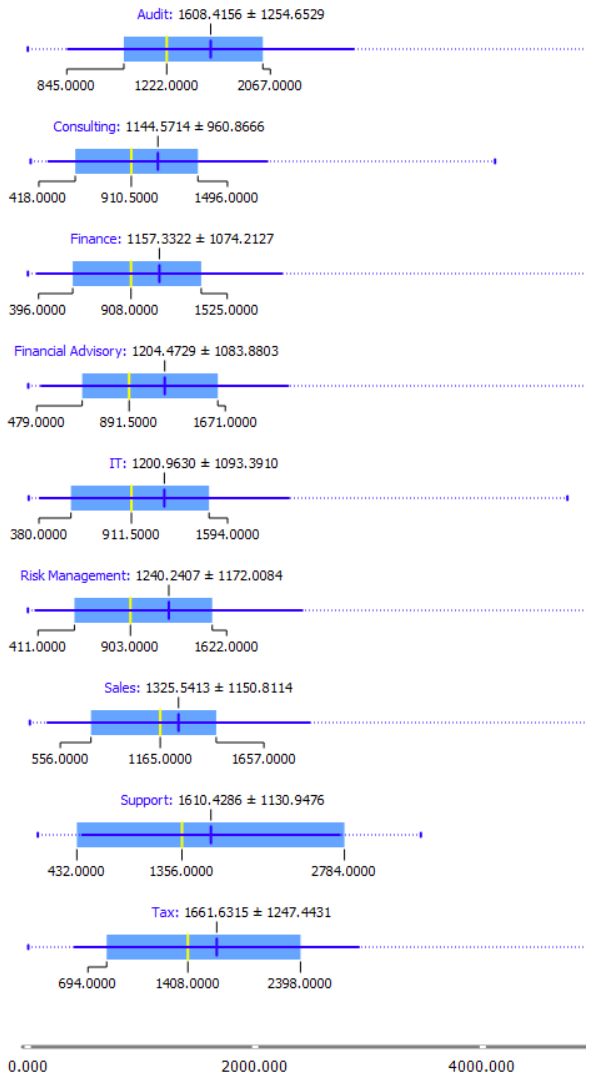


Terminated – Promoted

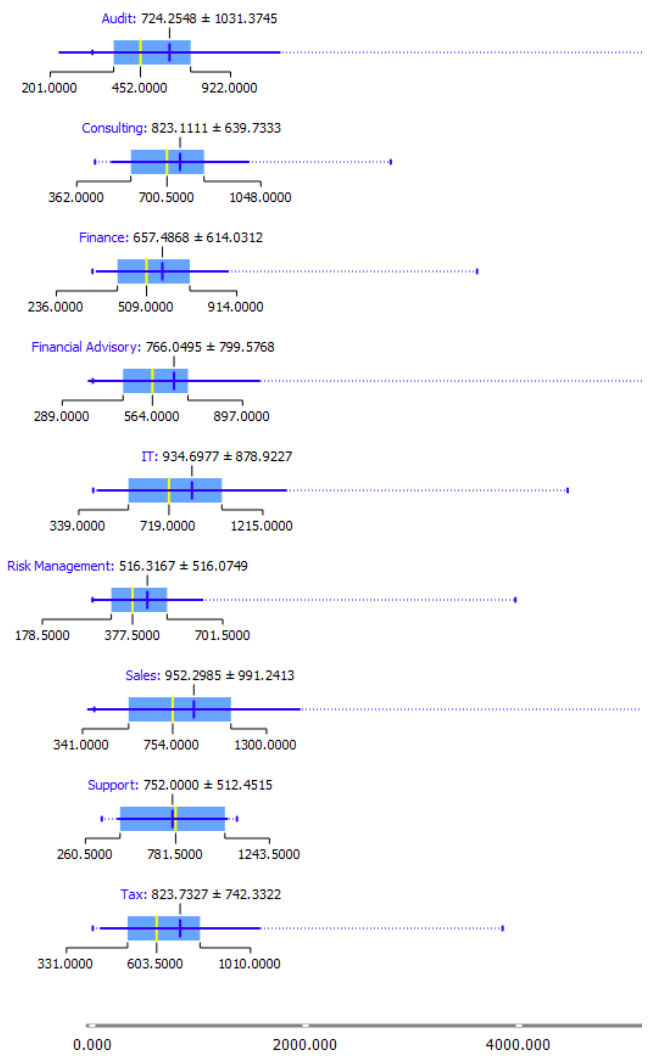


3.3 Department Thresholds

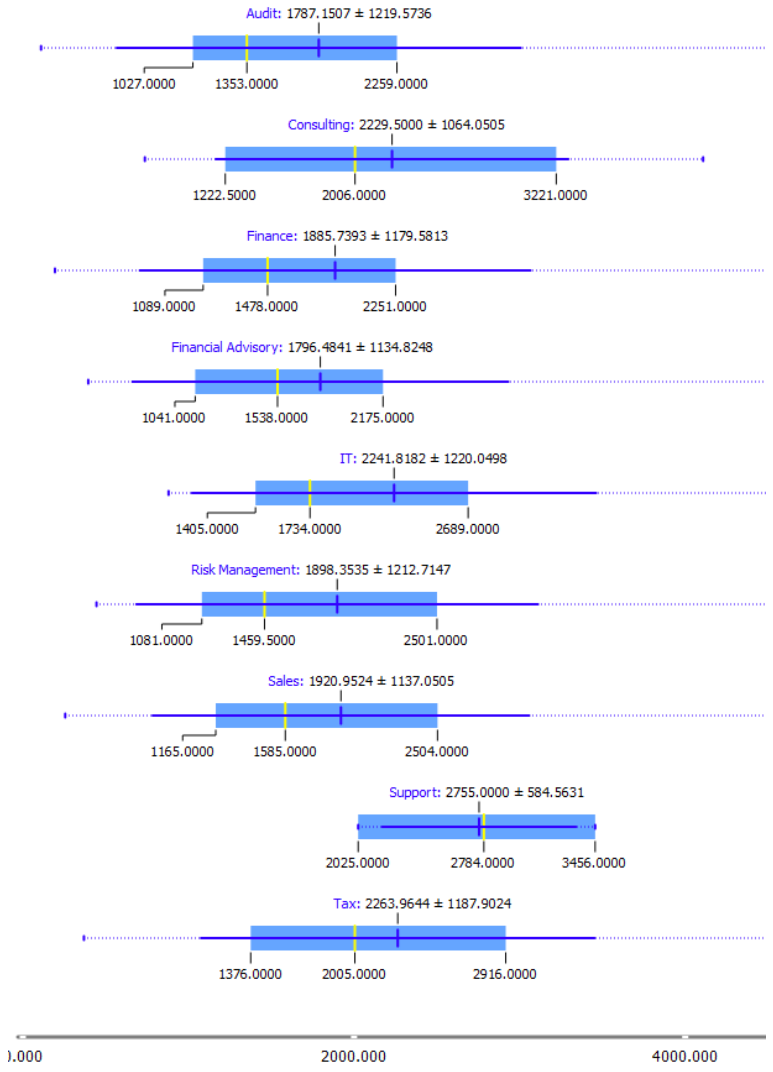
Terminated



Terminated – Not Promoted



Terminated – Promoted



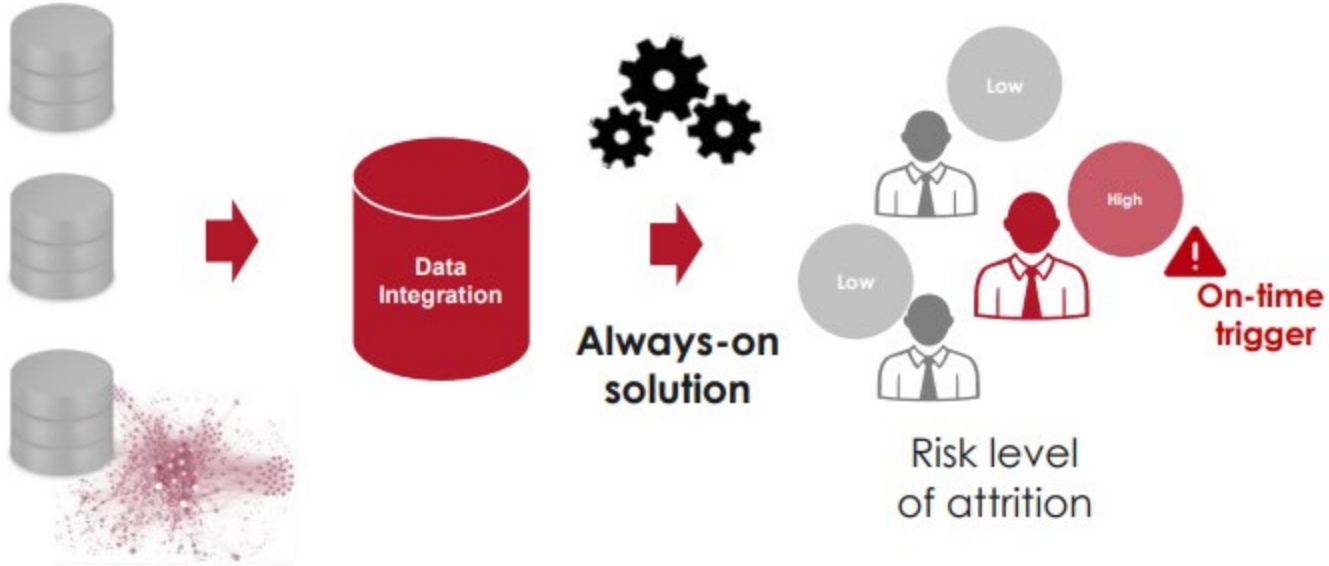
4. Risk Profiling Application Development

Web/standalone application can be developed where a HR Manager can understand attrition risk of each employee

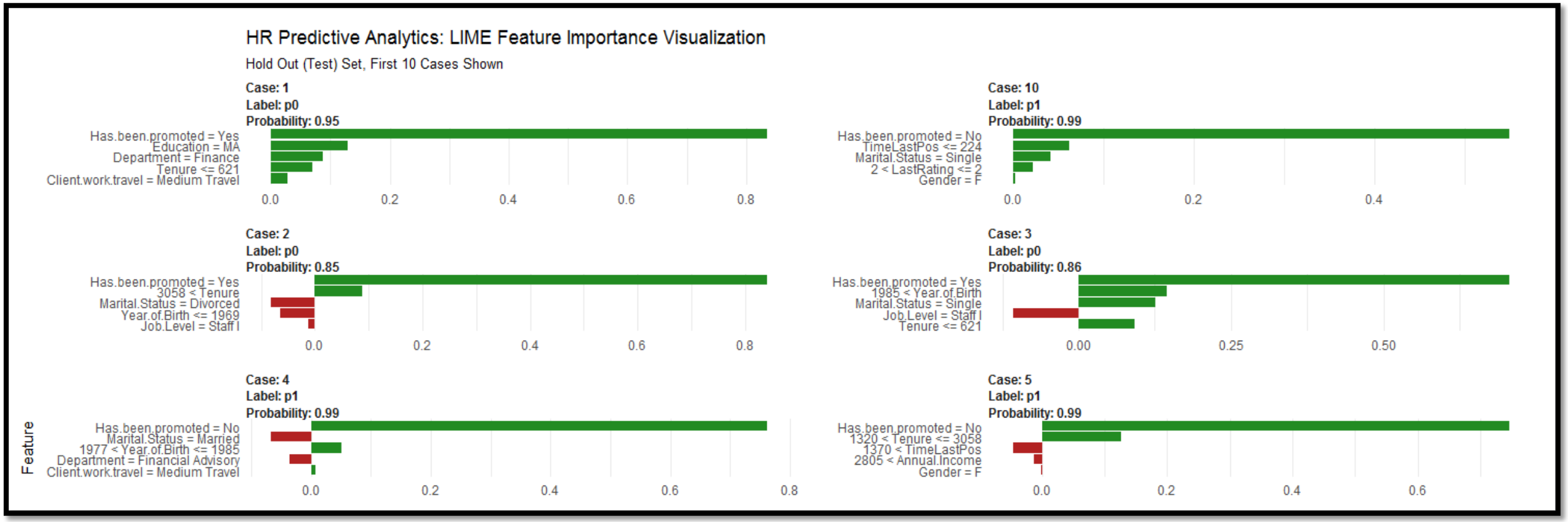
Integrating Data from several sources like Social Media , Behavioral ,Work related ,feedbacks and Peer to Peer reviews etc.

Focusing on modeling techniques we can classifier them into profiles and once model build can help us to understand the future attritions and features supporting and contradicting to particular choice employee made .

Can add Alerts and other more informative dashboards for HR managers to take necessary actions at right time with the help of relative thresholds.



4.1 Risk Profile : For each Individual : Model ready for deployment



4.2 Risk Profile : Matrix (Future Scope)

High	**Exceeds Expectations** Average APR 1 during the last 3 years, Can include once a 2 **Above Expectations** Average APR 2 during the last 3 years, Can include once a 3	HL 11 0.1%	HM 348 4.4%	HH 700 8.8%
Medium	**Meets Expectations** Average APR 3 during the last 3 years Can include once a 2 or a 4	ML 475 6.0%	MM 1971 24.8%	MH 1266 16.0%
Low	**Below Expectations** Average APR 4 or 5 during the last 3 years, can include once a 3	LL 139 1.8%	LM 210 2.6%	LH 51 0.64%
		Limited capacity, motivation or willingness to grow or ambition not demonstrated	Motivation and willingness to grow, capacity needs to be developed	Motivation, willingness and capacity to grow
		D (Low)	C (Medium)	A, B (High)

Can be created from more behavior Data to justify and to reduce the attrition of the employee

* Future Scope of work

Thank you

- Navaneesh Gangala