Assignment 7 - Employee Attrition

Business Case: In this case study we are analyzing the Employee Attrition for IBM HR Analytics Employee. High attrition rate causes issues at any organization. Not only does it affect the company financially but in terms of productivity / output of the work. It is crucial for any organization to understand the reasoning behind an employee's decision to leave to in-order to take preventive measures in the future. In our case we are analyzing 1470 employees for an organization to create a model which accurately analyses which employee might leave and provide them incentives to stay with the organization. In our case we are checking the same based on employee demographic, and job-related data such as job satisfaction, working hours, overtime, hourly rate, job level, etc and then predicting attrition. This model can be used by HR team of an organization to predict the employees who might leave and provide incentives to retain them.

Question 1 and 2:

Below steps were performed for data preprocessing:

1. The field "Employee Number" is discarded due to high cardinality.
2. The field "Employee Count" is discarded as it only had one value(that is '1') for all the entries and hence made no impact on prediction.
3. The field "Over 18" is discarded as it only had one value(that is 'y') for all the entries and hence made no impact on prediction.
4. The field "Standard hour" is discarded as it only had one value(that is '80 hrs') for all the entries and hence made no impact on prediction.
5. "Job Level " is converted into categorical feature as it has 1-5 numeric values.
6. "Stock option level" is converted into categorical feature as it has 0-3 numeric values.
7. Converted "Education" , "EnvironmentSatisfaction" , "JobInvolvement" , "JobSatisfaction", "PerformanceRating", "RelationshipSatisfaction"," WorkLifeBalance" to categorical features as per the given values:

IBM HR Analytics Employee Attrition & Performance

Education

1 'Below College'

2 'College'

3 'Bachelor'

4 'Master'

5 'Doctor'

EnvironmentSatisfaction

1 'Low'

2 'Medium'

3 'High'

4 'Very High'

JobInvolvement

1 'Low'

2 'Medium'

3 'High'

4 'Very High'

IBM HR Analytics Employee Attrition & Performance

JobSatisfaction

1 'Low'

2 'Medium'

3 'High'

4 'Very High'

PerformanceRating

1 'Low'

2 'Good'

3 'Excellent'

4 'Outstanding'

RelationshipSatisfaction

1 'Low'

2 'Medium'

3 'High'

4 'Very High'

WorkLifeBalance

1 'Bad'

IBM HR Analytics Employee Attrition & Performance

PerformanceRating

1 'Low'

2 'Good'

3 'Excellent'

4 'Outstanding'

RelationshipSatisfaction

1 'Low'

2 'Medium'

3 'High'

4 'Very High'

WorkLifeBalance

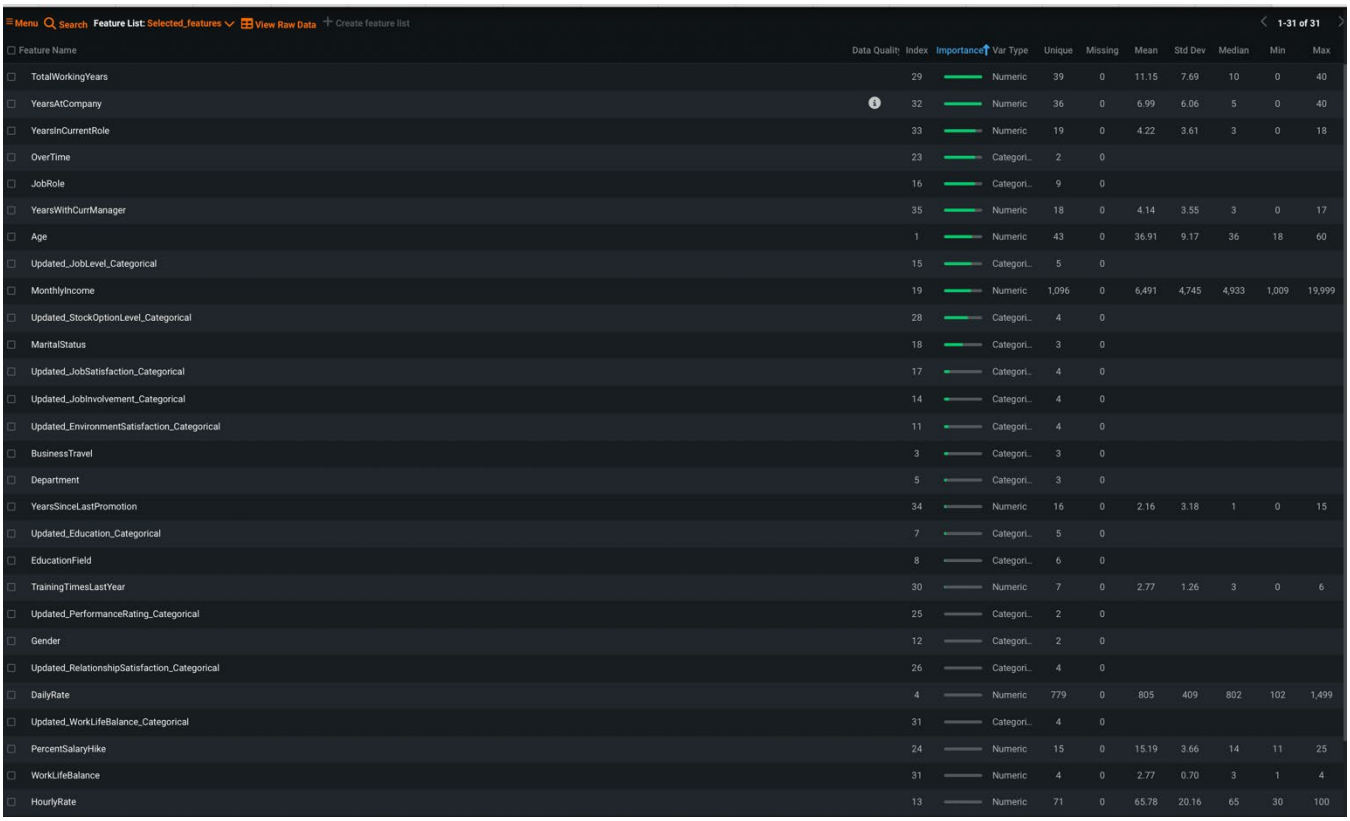
1 'Bad'

2 'Good'

3 'Better'

4 'Best'

Final selected features:


Menu Search Feature List: Selected_features View Raw Data Create feature list 1-31 of 31

Feature Name	Data Quality	Index	Importance	Var Type	Unique	Missing	Mean	Std Dev	Median	Min	Max
TotalWorkingYears		29		Numeric	39	0	11.15	7.69	10	0	40
YearsAtCompany		32		Numeric	36	0	6.99	6.06	5	0	40
YearsInCurrentRole		33		Numeric	19	0	4.22	3.61	3	0	18
OverTime		23		Categori..	2	0					
JobRole		16		Categori..	9	0					
YearsWithCurrManager		35		Numeric	18	0	4.14	3.55	3	0	17
Age		1		Numeric	43	0	36.91	9.17	36	18	60
Updated_JobLevel_Categorical		15		Categori..	5	0					
MonthlyIncome		19		Numeric	1,096	0	6,491	4,745	4,933	1,009	19,999
Updated_StockOptionLevel_Categorical		28		Categori..	4	0					
MaritalStatus		18		Categori..	3	0					
Updated_JobSatisfaction_Categorical		17		Categori..	4	0					
Updated_JobInvolvement_Categorical		14		Categori..	4	0					
Updated_EnvironmentSatisfaction_Categorical		11		Categori..	4	0					
BusinessTravel		3		Categori..	3	0					
Department		5		Categori..	3	0					
YearsSinceLastPromotion		34		Numeric	16	0	2.16	3.18	1	0	15
Updated_Education_Categorical		7		Categori..	5	0					
EducationField		8		Categori..	6	0					
TrainingTimesLastYear		30		Numeric	7	0	2.77	1.26	3	0	6
Updated_PerformanceRating_Categorical		25		Categori..	2	0					
Gender		12		Categori..	2	0					
Updated_RelationshipSatisfaction_Categorical		26		Categori..	4	0					
DailyRate		4		Numeric	779	0	805	409	802	102	1,499
Updated_WorkLifeBalance_Categorical		31		Categori..	4	0					
PercentSalaryHike		24		Numeric	15	0	15.19	3.66	14	11	25
WorkLifeBalance		31		Numeric	4	0	2.77	0.70	3	1	4
HourlyRate		13		Numeric	71	0	65.78	20.16	65	30	100

Feature Name	Data Quality	Index	Importance	Var Type	Unique	Missing	Mean	Std Dev	Median	Min	Max
OverTime		23		Categori..	2	0					
JobRole		16		Categori..	9	0					
YearsWithCurrManager		35		Numeric	18	0	4.14	3.55	3	0	17
Age		1		Numeric	43	0	36.91	9.17	36	18	60
Updated_JobLevel_Categorical		15		Categori..	5	0					
MonthlyIncome		19		Numeric	1,096	0	6,491	4,745	4,933	1,009	19,999
Updated_StockOptionLevel_Categorical		28		Categori..	4	0					
MaritalStatus		18		Categori..	3	0					
Updated_JobSatisfaction_Categorical		17		Categori..	4	0					
Updated_JobInvolvement_Categorical		14		Categori..	4	0					
Updated_EnvironmentSatisfaction_Categorical		11		Categori..	4	0					
BusinessTravel		3		Categori..	3	0					
Department		5		Categori..	3	0					
YearsSinceLastPromotion		34		Numeric	16	0	2.16	3.18	1	0	15
Updated_Education_Categorical		7		Categori..	5	0					
EducationField		8		Categori..	6	0					
TrainingTimesLastYear		30		Numeric	7	0	2.77	1.26	3	0	6
Updated_PerformanceRating_Categorical		25		Categori..	2	0					
Gender		12		Categori..	2	0					
Updated_RelationshipSatisfaction_Categorical		26		Categori..	4	0					
DailyRate		4		Numeric	779	0	805	409	802	102	1,499
Updated_WorkLifeBalance_Categorical		31		Categori..	4	0					
PercentSalaryHike		24		Numeric	15	0	15.19	3.66	14	11	25
WorkLifeBalance		31		Numeric	4	0	2.77	0.70	3	1	4
HourlyRate		13		Numeric	71	0	65.78	20.16	65	30	100
MonthlyRate		20		Numeric	1,156	0	14,316	7,115	14,224	2,094	26,999
NumCompaniesWorked		21		Numeric	10	0	2.72	2.55	2	0	9
DistanceFromHome		6		Numeric	29	0	9.23	8.10	7	1	29

Question 3:

Business Case: In this case study we are analyzing the Employee Attrition for IBM HR Analytics Employee. High attrition rate causes issues at any organization. Not only does it affect the company financially but in terms of productivity / output of the work. It is crucial for any organization to understand the reasoning behind an employee's decision to leave to in-order to take preventive measures in the future. In our case we are analyzing 1470 employees for an organization to create a model which accurately analyses which employee might leave and provide them incentives to stay with the organization. In our case we are checking the same based on employee demographic, and job-related data such as job satisfaction, working hours, overtime, hourly rate, job level, etc and then predicting attrition. This model can be used by HR team of an organization to predict the employees who might leave and provide incentives to retain them.

Payoff Matrix:

Assumptions:

1. The average salary of an employee is $150,000.
2. In our case we are assuming that the employee decides to stay with the company after a raise is offered. The company offers the employee raise of $20,000.
3.If an employee leaves the company the same position is filled by a new employee for $180,000.
4. The hiring process costs $1000 including all the miscellaneous costs.
5.The training cost is $800 for each new employee.

True Positive: The model accurately predicts that the employee is going to leave the company and therefore the company needs to bear the bonus offered to retain the employee.

TP: -$20,000

True Negative: The model accurately identifies the employee who is not going to switch the company and therefore are not offered any bonus.

TN: $0

False Positive: The model inaccurately identifies an employee as someone who is going to leave the organization and is offered a bonus although in reality the employee was not going to leave.

FP: -$20,000

False Negative: The model inaccurately identifies an employee as someone who is not going to leave the organization and therefore is not offered any bonus and the employee eventually leaves.

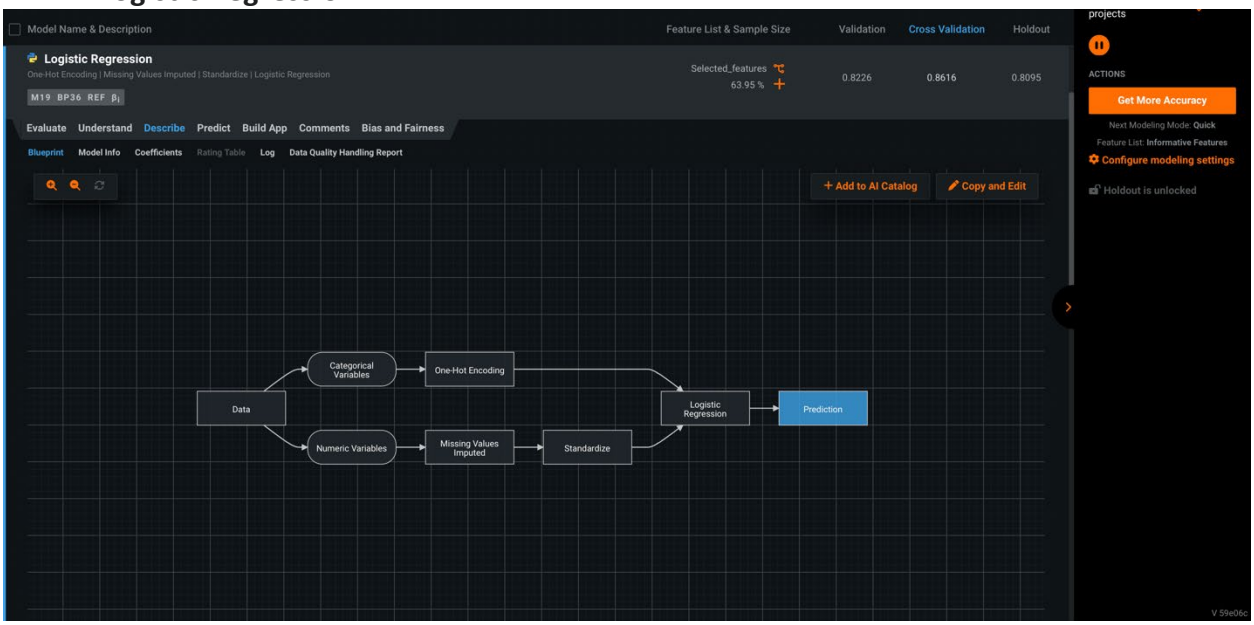
FN: -$30,000(difference between the salary of the new hire) - $1000(hiring cost)-$800(training cost)
= -$31,800

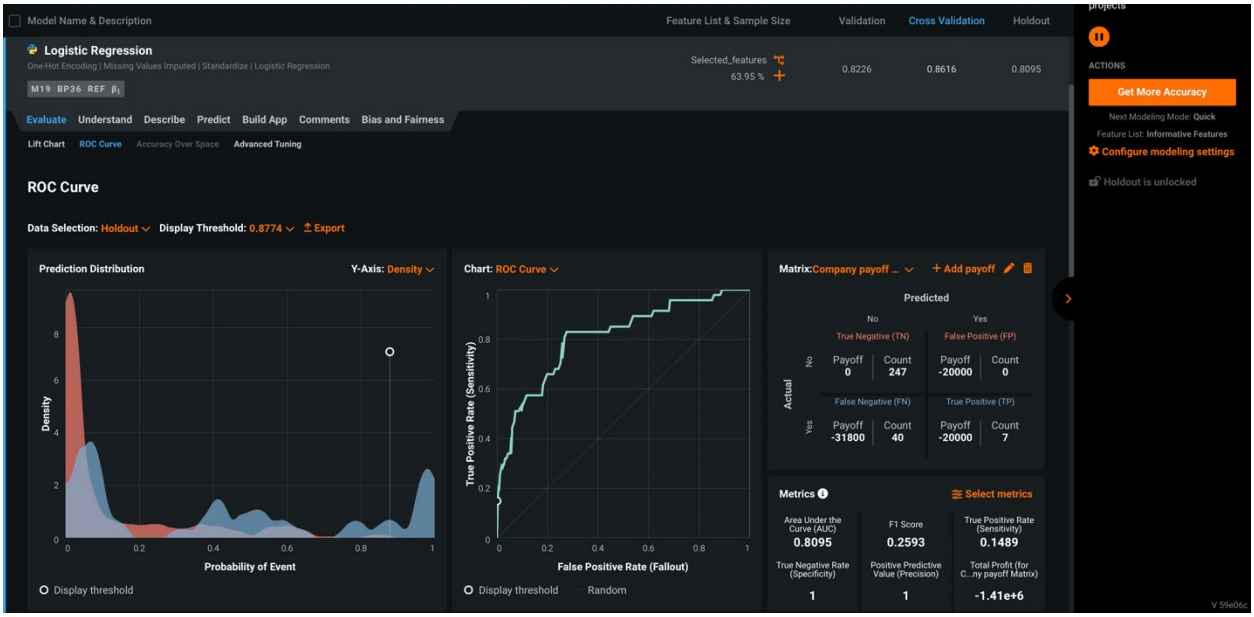
	Predicted 0	Predicted 1
Actual 0	$0	-$20,000
Actual 1	-$31,800	-$20,000

Best Metric to evaluate the model in our case is the Maximum payoff, as in our case our end goal is to minimize the cost for the company. Maximum payoff metric assigns costs and benefits to different types of correct and incorrect predictions (true positives/true negatives and false positives/false negatives) and help evaluate the required profit/losses based on the given case. We need a profit metric to evaluate the same and hence maximum payoff is ideal for this case.

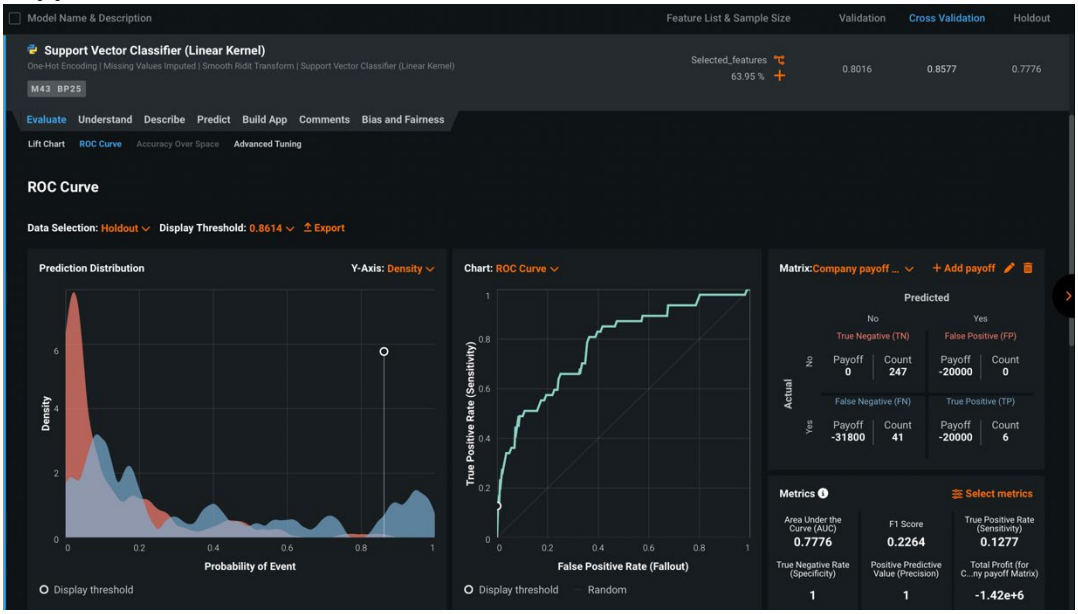
Question 4:

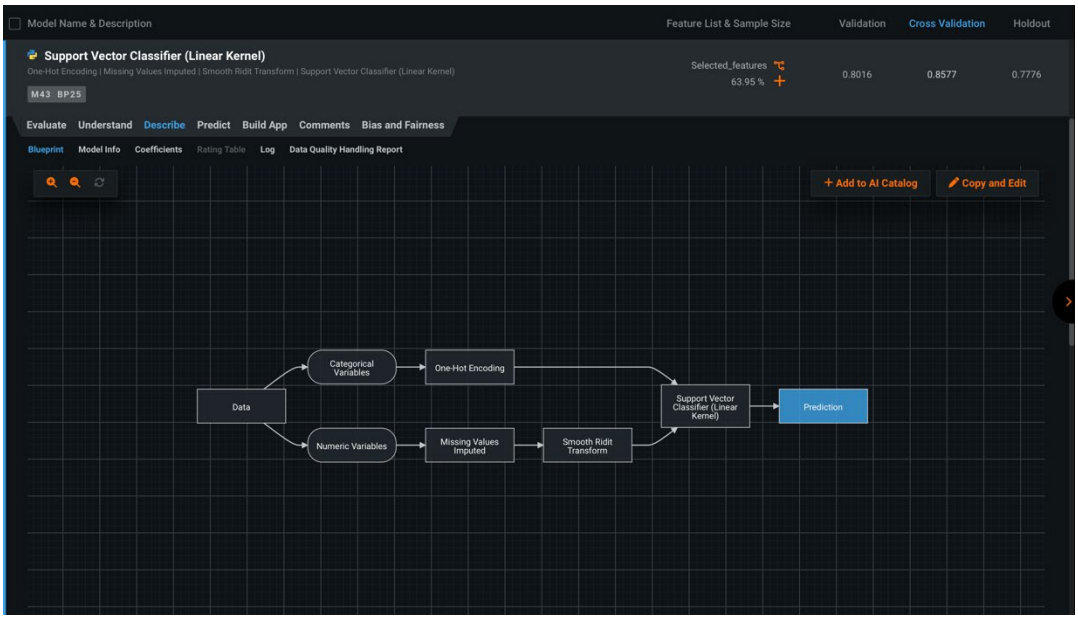
1. **Logistic Regression :**



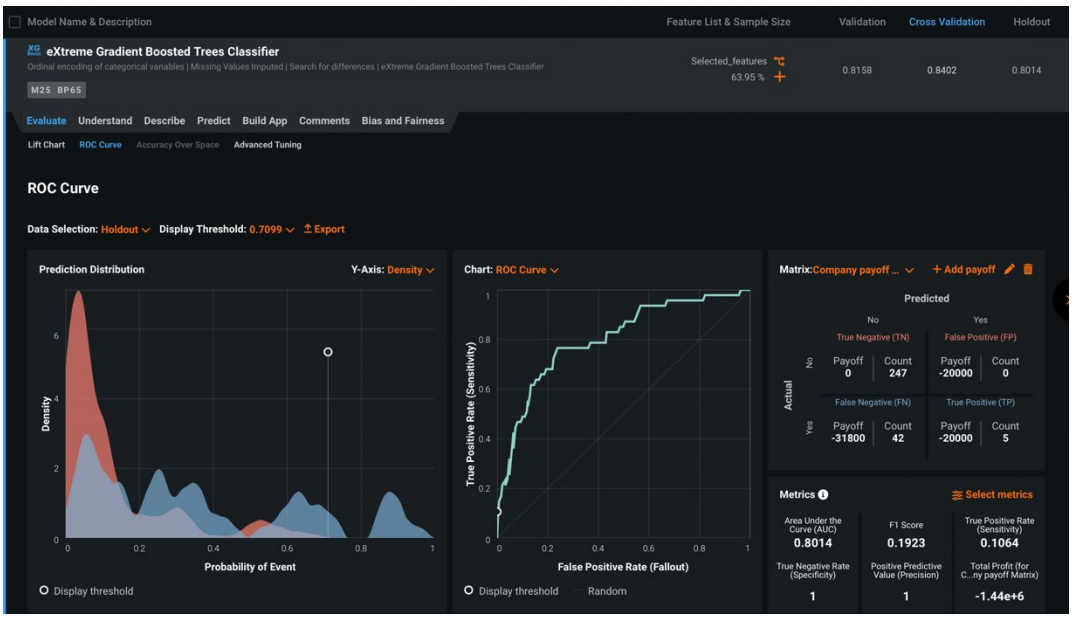


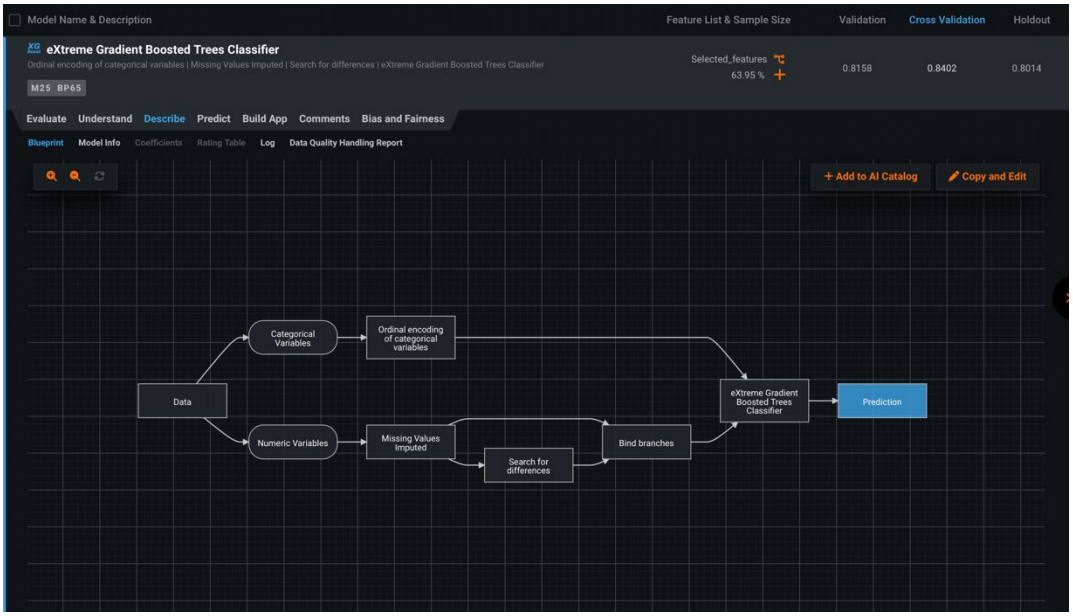
2. Support Vector Machine



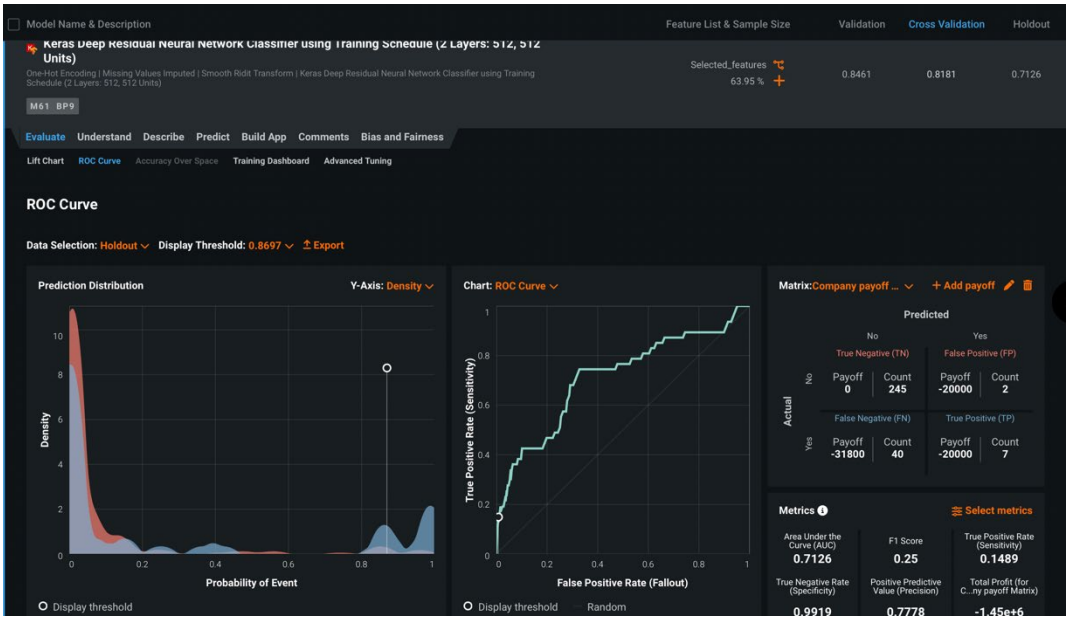


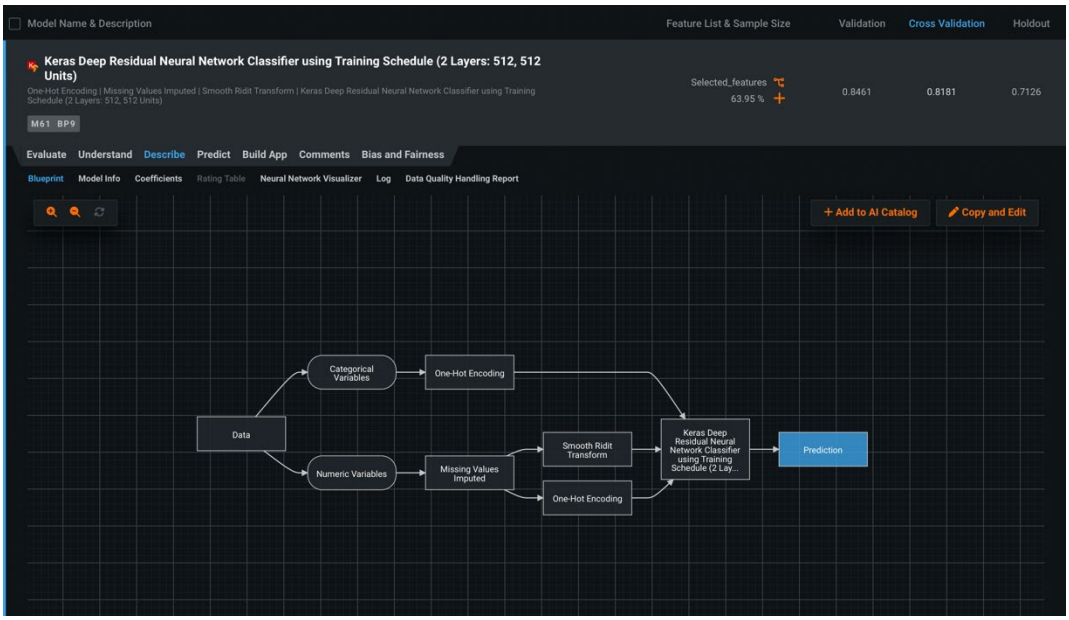
3. Gradient boosted Tree:



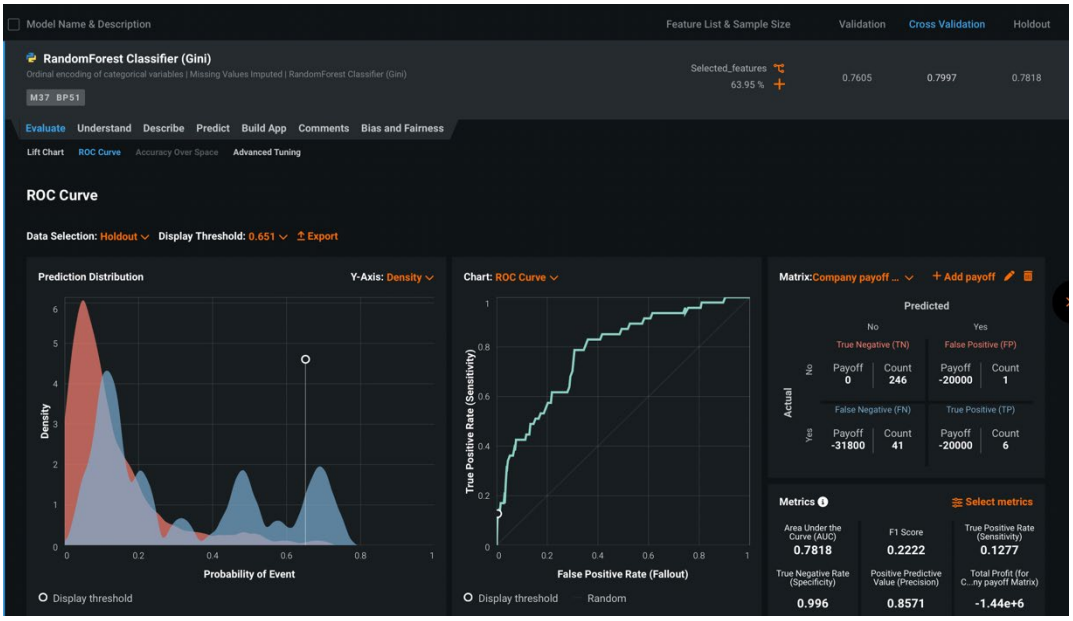


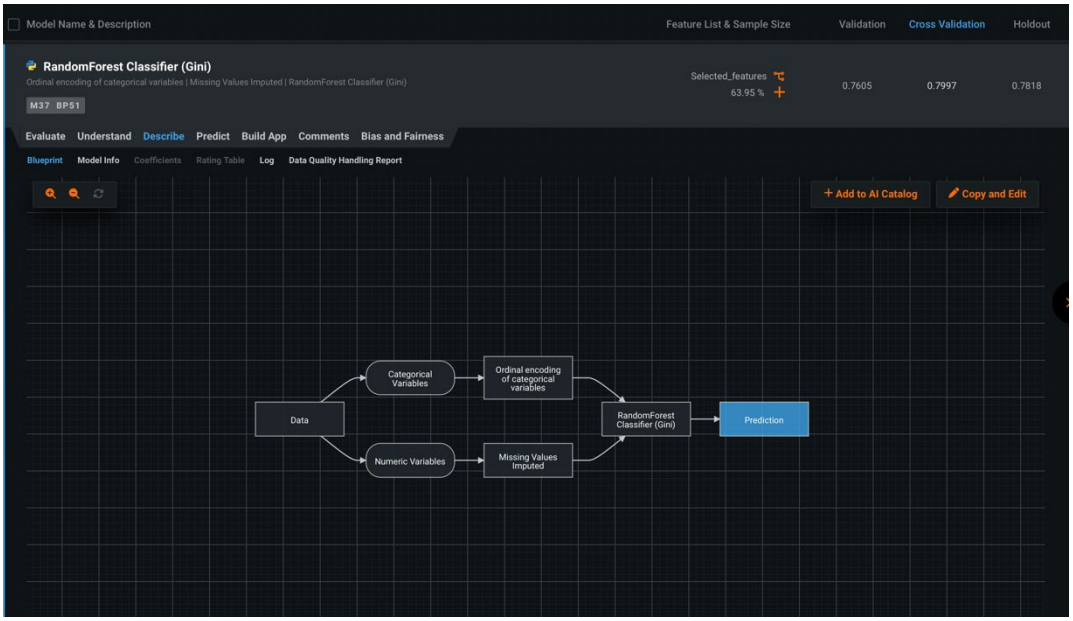
4. Neural Network



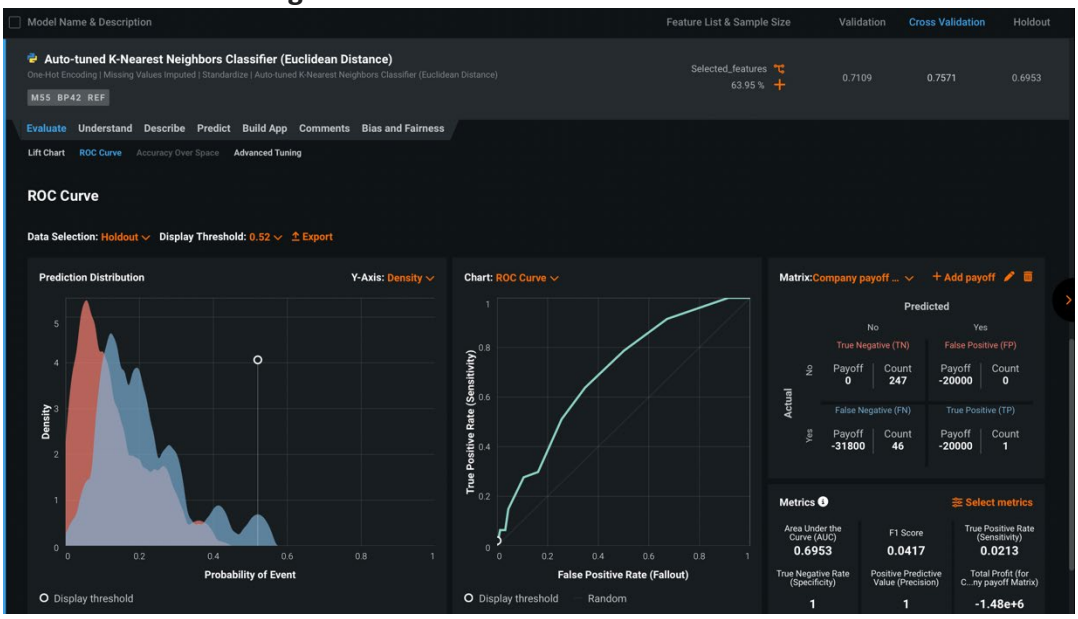


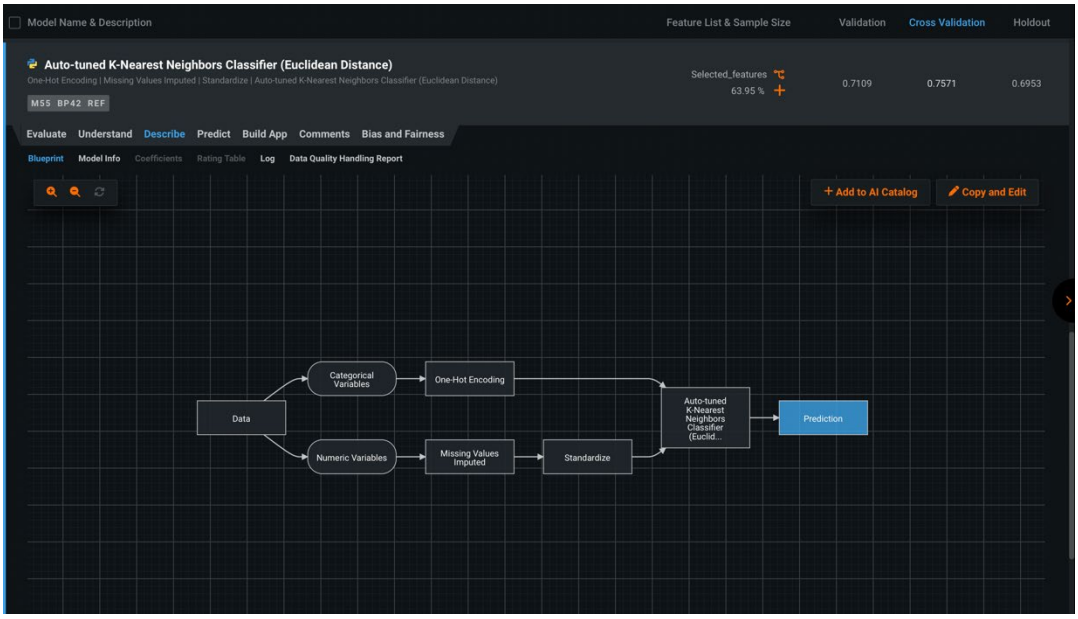
5. Random forest Classifier:



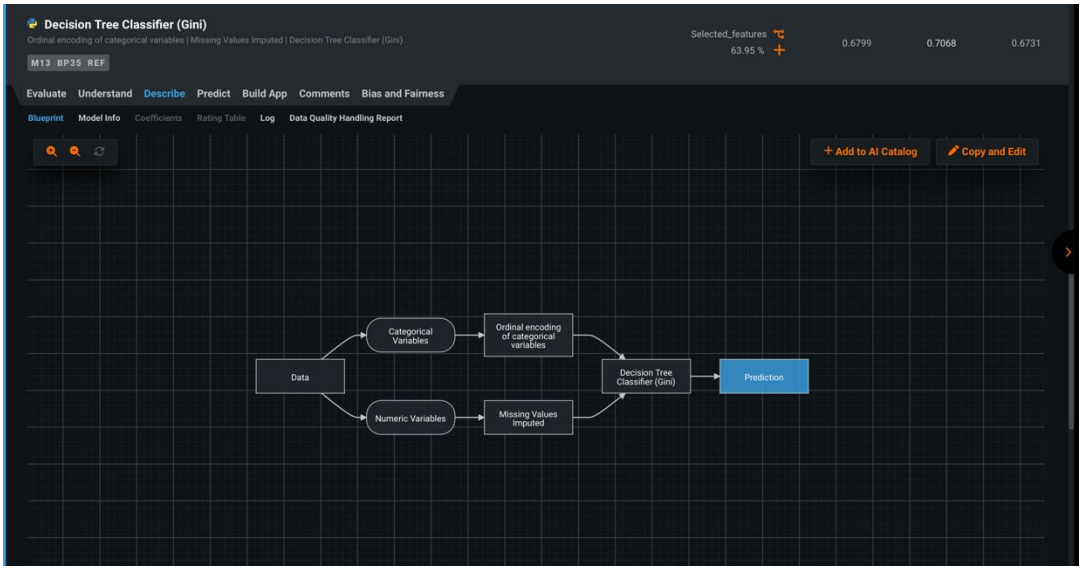


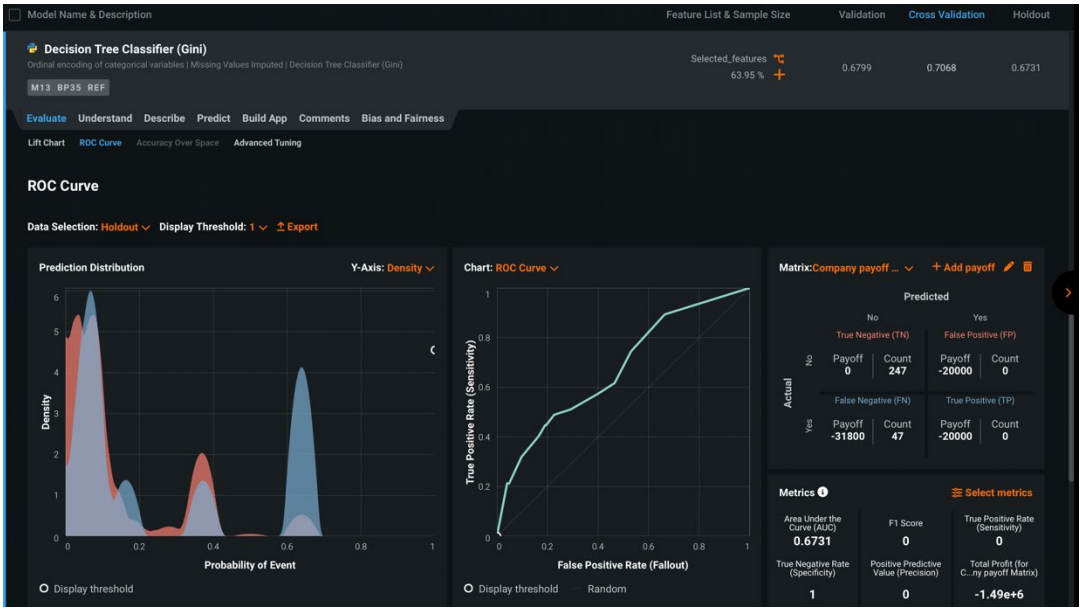
6. K-Nearest Neighbors:





7. Decision Tree:



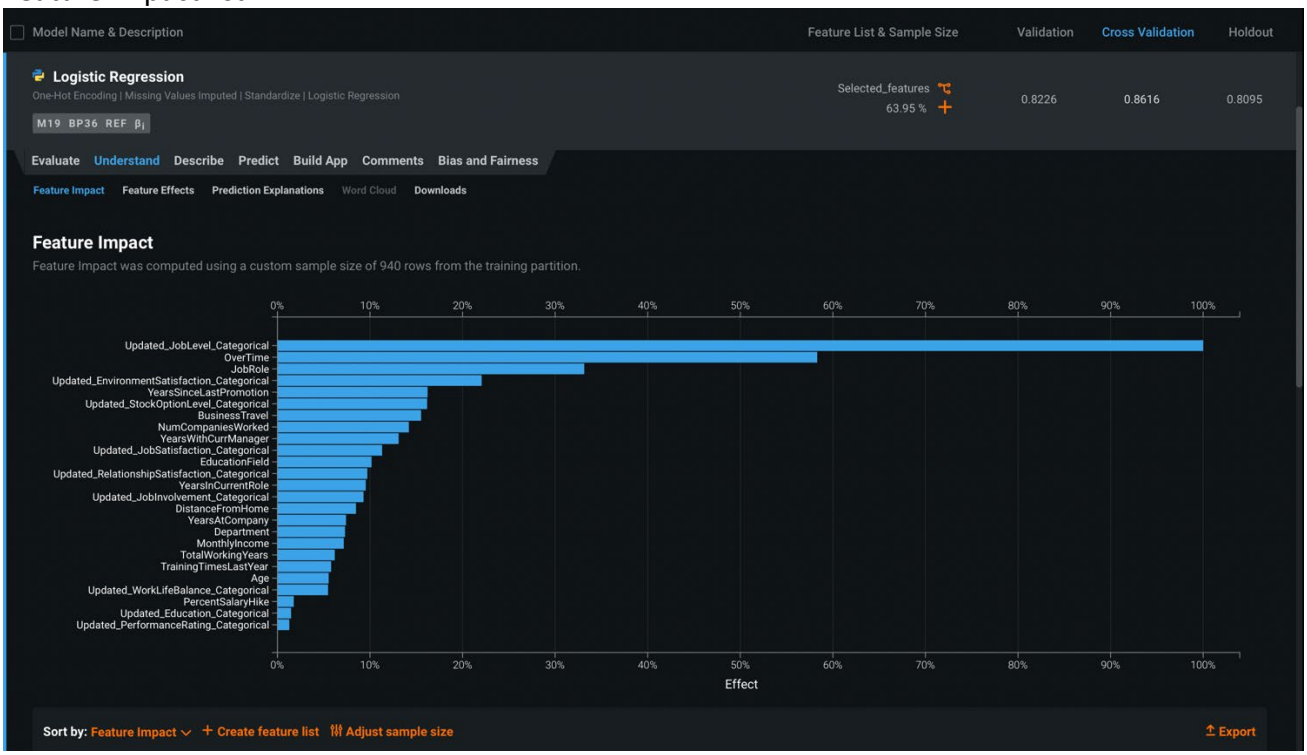


Model	AUC	F1	Recall	Precision	Specificity	Maximum Pay off
Decision Tree Classifier (Gini)(Threshold: 1)	0.6731	0	0	0	1	-$1,490,000
Logistic Regression(Threshold: 0.8774)	0.8095	0.2593	0.1489	1	1	-$1,410,000
Support Vector Classifier (Linear Kernel)(Threshold: 0.8614)	0.7776	0.2264	0.1277	1	1	-$1,420,000
Gradient Boosted Trees Classifier(Threshold: 0.7099)	0.8014	0.1923	0.1064	1	1	-$1,440,000
Keras Deep Residual Neural Network Classifier using Training Schedule (2 Layers: 512, 512 Units) (Threshold: 0.8697)	0.7126	0.25	0.1489	0.7778	0.9919	-$1,450,000
RandomForest Classifier (Gini)(Threshold: 0.651)	0.7818	0.2222	0.1277	0.8571	0.996	-$1,440,000
Auto-tuned K-Nearest Neighbors Classifier (Euclidean Distance)(Threshold:0.52)	0.6953	0.0417	0.0213	1	1	-$1,480,000

The best performing model is Logistic Regression. It provides the payoff metric of -$1,410,000. Best Metric to evaluate the model in our case is the Maximum payoff, as in our case our end goal is to minimize the cost for the company. Maximum payoff metric assigns costs and benefits to different types of correct and incorrect predictions (true positives/true negatives and false positives/false negatives) and help evaluate the required profit/losses based on the given case. We need a profit metric to evaluate the same and hence maximum payoff is ideal for this case.

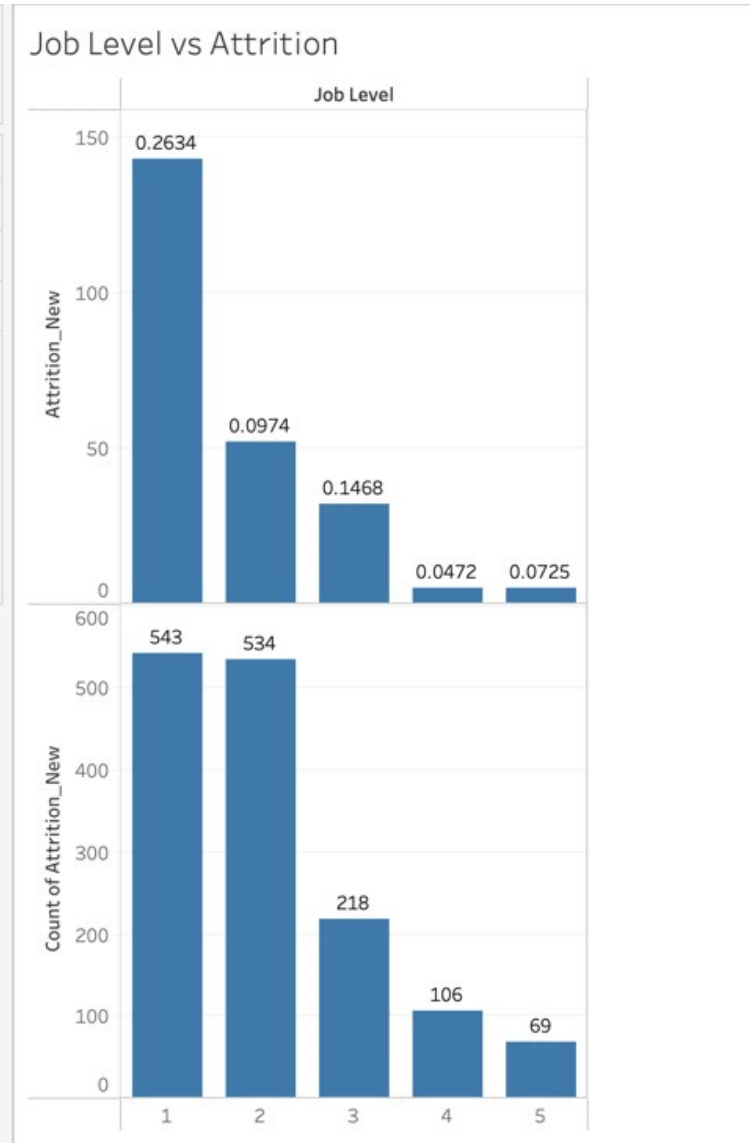
Question 5:

Feature impact list :



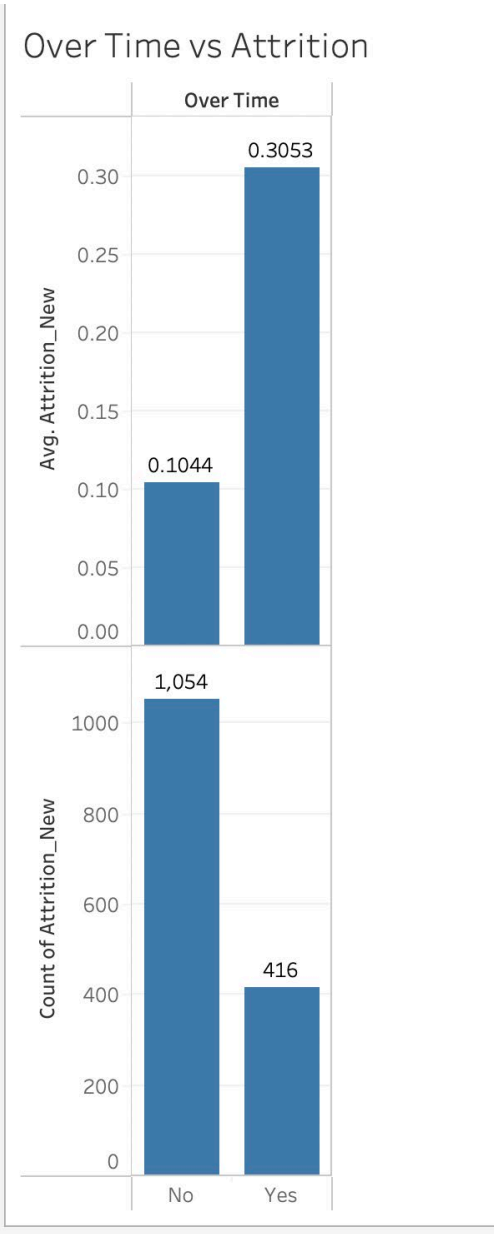
Visualization of the top 5 features:

1. Job Level



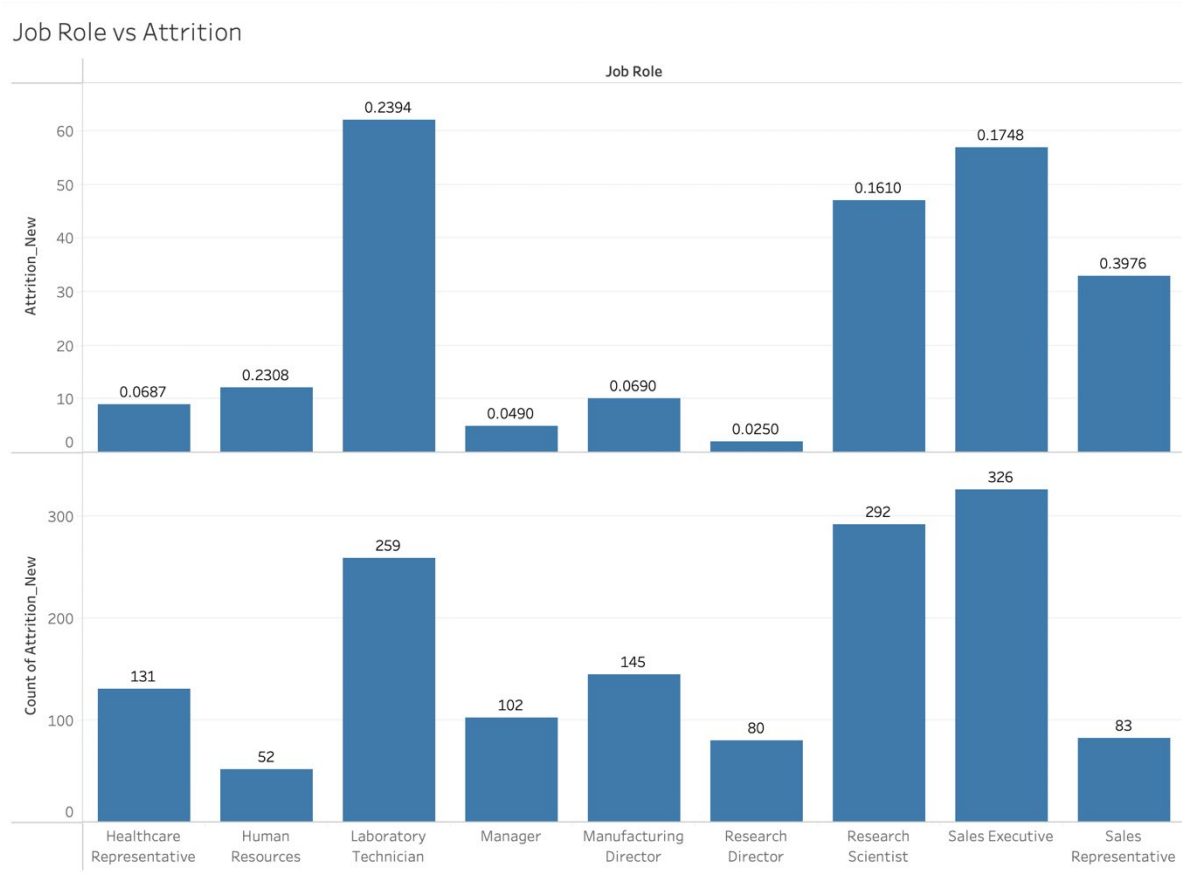
Highest attrition is observed at job level 1. For a count of 543 employees the attrition is around 26.34%. Job level 2 has almost similar employee count of 534 but as seen the attrition rate is 9.74% which is 2.7 times less than job level 1. Which indicates people at level 1 are more likely to leave.

2. Over Time



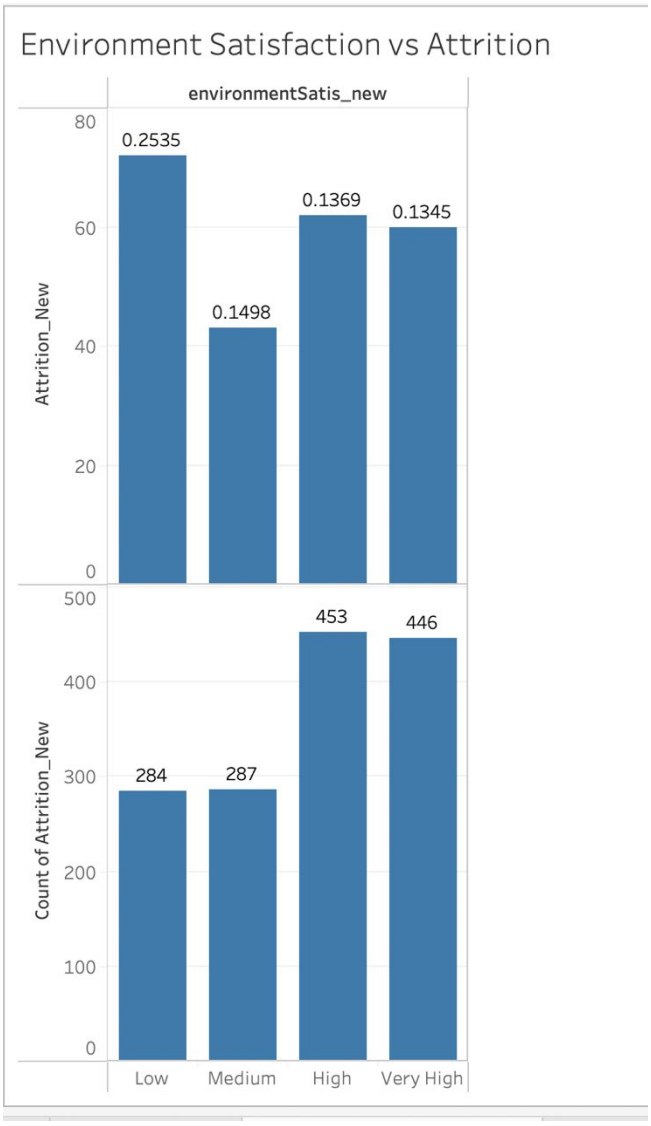
As observed employee who are not working overtime are more in count (1054 in our case) however they have a lower attrition rate of 10.44% when compared to employee who work over time (416 count in our case) with 30.53% attrition rate. Employees who work overtime have a 3 times attrition rate compared to employees who do not work overtime.

3. Job Role



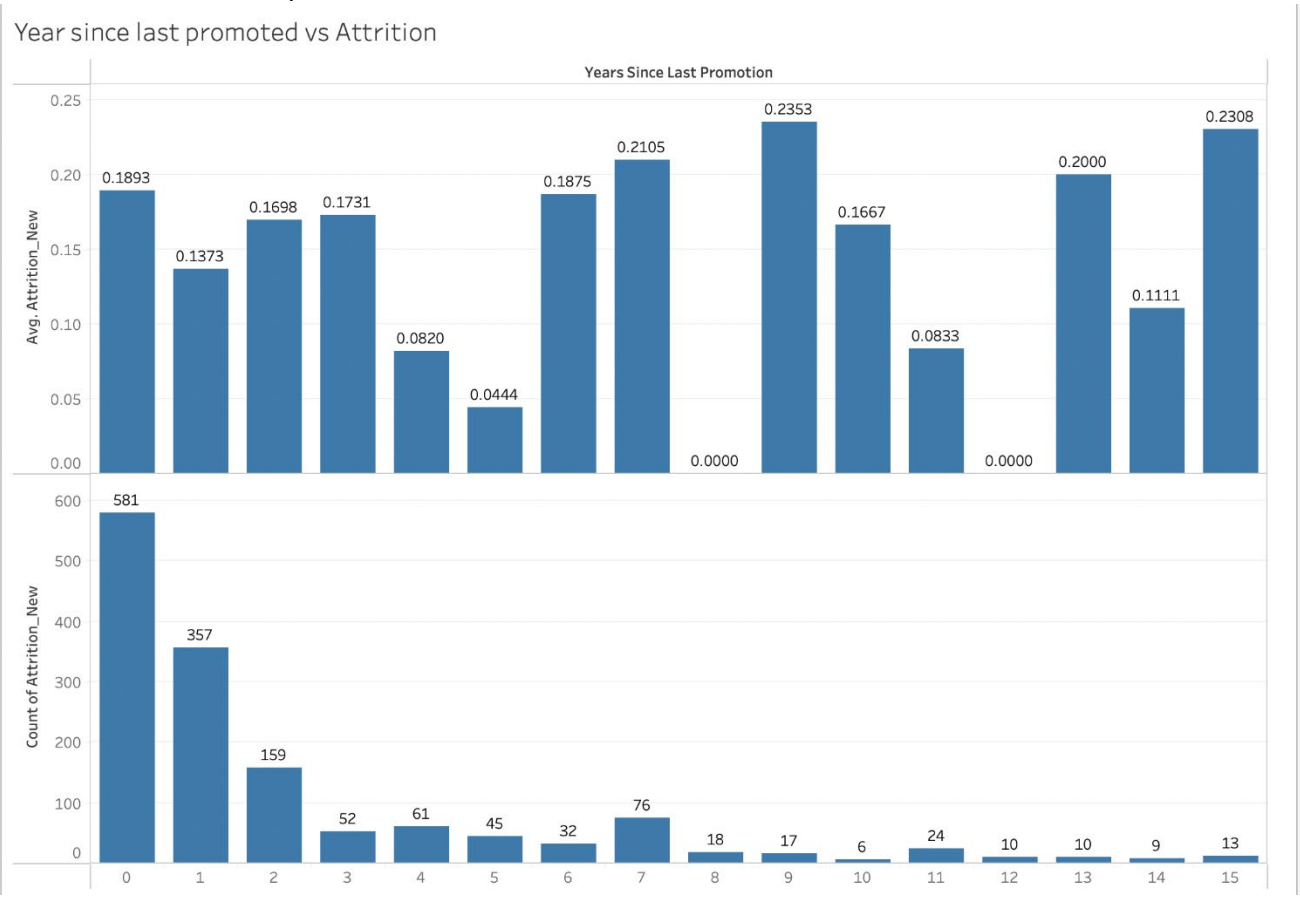
As observed Sales representatives have the highest attrition rate of 39.76% , followed by Laboratory technicians at 23.94% , followed by Human Resources at 23.08% , followed by Sales executive at 17.48%, followed by Research Scientist at 16.10%. Others have attrition rate below 7%.

4. Environment Satisfaction



As observed employee who have “lower” environmental satisfaction (in our case with count 284) have the highest attrition rate of 25.35%. The attrition rate is almost double for these employees when compared to those who have “high” (13.69% attrition with count 453) or “very high” (13.45% attritionwith count 446) environmental satisfaction.

5. Year since last promoted



As observed the attrition is highest for employee who has not been promoted since last 9 years for a total of 23.53% closely followed by employees who have not been promoted since last 15 years for a total of 23.08%.

Question 6:

Observed Effect	Recommendation
<u>Feature: Job Level</u> Highest attrition is observed at job level 1. For a count of 543 employees the attrition is around 26.34%. Job level 2 has almost similar employee count of 534 but as seen the attrition rate is 9.74% which is 2.7 times less than job level 1. Which indicates people at level 1 are more likely to leave.	The data dictionary does not define what the job level exactly indicates. However, one can observe that the attrition is higher at job level 1 when compared to job level 2 even though the employee count is almost similar. Therefore, the HR team should consult both these groups to check what might be an issue at job level 1 when compared to 2.
<u>Feature: Over Time</u> As observed employee who are not working overtime are more in count (1054 in our case) however they have a lower attrition rate of 10.44% when compared to employee who work over time (416 count in our case) with 30.53% attrition rate. Employees who work overtime have a 3 times attrition rate compared to employees who do not work overtime.	As the data indicates that employees working overtime have higher attrition rate. Therefore, the HR should connect with the required team managers to device solution for the overtime problem. They could also increase the pay rate for overtime work and even bring in more people to join the required team in-order to reduce the burden over current employees and reduce overtime duration for such employees.
<u>Feature: Job Role</u> As observed Sales representatives have the highest attrition rate of 39.76% , followed by Laboratory technicians at 23.94% , followed by Human Resources at 23.08% , followed by Sales executive at 17.48%, followed by Research Scientist at 16.10%. Others have attrition rate below 7%.	The HR should consult the employees from specific roles with high attrition rate to check where the issues are and address the same. If need be, they can involve subject matter experts to understand and device solution for the problems observed .As this will help to reduce the attrition rate.
<u>Feature: Environment Satisfaction</u> As observed employee who have “lower” environmental satisfaction (in our case with count 284) have the highest attrition rate of 25.35%. The attrition rate is almost double for these employees when compared to those who have “high” (13.69% attrition with count 453) or “very high” (13.45% attritionwith count 446) environmental satisfaction.	The HR should work with people who have “lower” environmental satisfaction to understand where the issue is. This may be related to problems such as to include tea-time connect within teams during work hours, or to buy faster high processing devices. Once the HR knows what can help to increase the environmental satisfaction for these employees the attrition rate can be reduced.
<u>Feature: Year since last promoted</u>	The attrition rate is relatively lower in employees who have been recently promoted

As observed the attrition is highest for employee who has not been promoted since last 9 years for a total of 23.53% closely followed by employees who have not been promoted since last 15 years for a total of 23.08%.	when compared to those who have not been promoted for a very long time. The HR should reach out to relevant team mangers and understand as to why these employees are not getting promoted. If there are deserving employees as per their performance the HR should assist with the required due promotions to lower the attrition rate of such employees.