Project Report

Employee Attrition Prediction

Group 6

Members:

Ruchak Vira

Soham Pendse

Priyal Bamnikar

Table Of Contents

summary	satisfaction_level	last_evaluation	number_project	time_spend_company	average_monthly_hours
count	14999	14999	14999	14999	14999
mean	0.6128335220371535	0.7161017401400102	3.80305353690246	3.498233215547703	201.0503366891126
stddev	0.24863065162689157	0.17116911257197398	1.2325923553183513	1.4601362305354808	49.94309937128406
min	0.09	0.36	2	2.0	96.0
max	1.0	1.0	7	10.0	310.0

Abstract

Overview

Employee attrition is a ratio of employees who leave a company & the employees who do not. This predictive study can be used by a company for strategizing to reduce attrition. By identifying the primary causes of the high attrition rate, a company can improve the working environment for its employees. Furthermore, we are also trying to examine factors influencing job satisfaction levels as these may be a key factor in improving employee retention strategies for an organization.

Dataset Description

The dataset is structured to account for factors impacting employee attrition. It consists of integer, float, string, and categorical data. Factors like satisfaction level, salary, age, evaluation score, department and average monthly hours are part of the dataset.

Number of Rows	14999
Number of Columns	10

Predictors

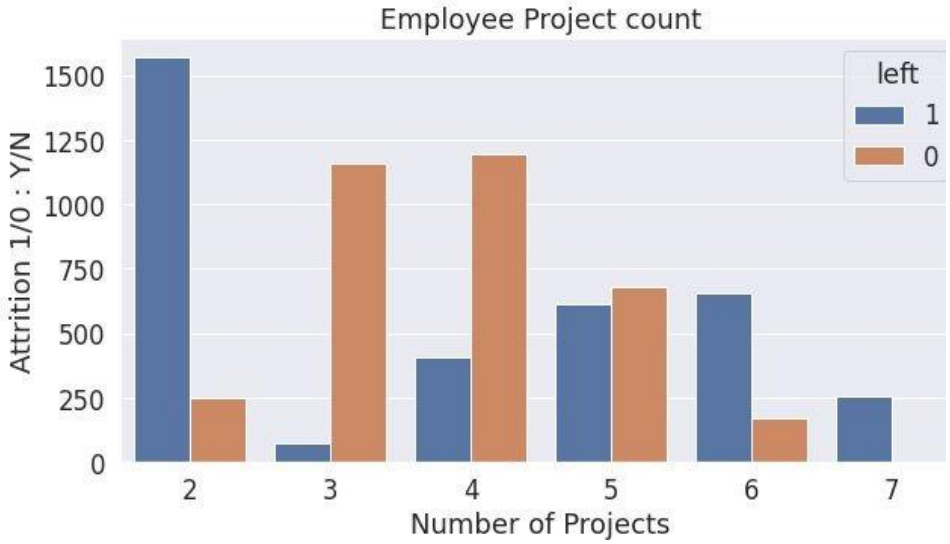
Satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	promotion_last_5years	department	salary

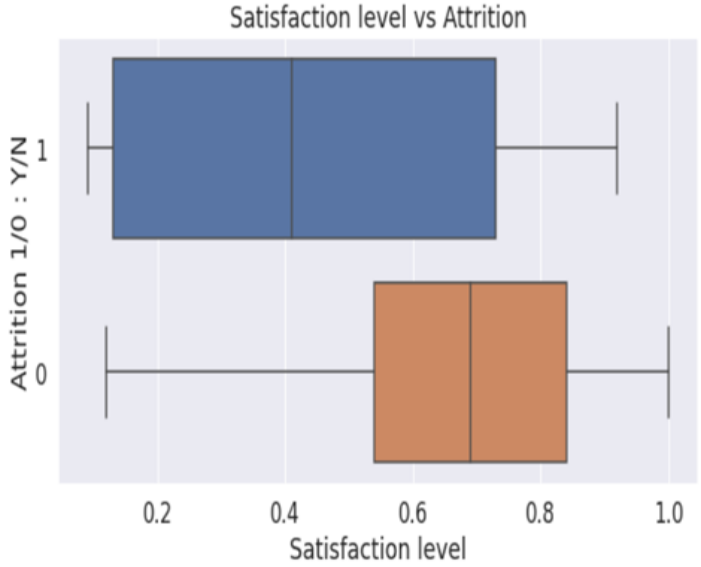
Link to Dataset

https://raw.githubusercontent.com/rnvira/Employee-Attrition/main/Employee_Attrition.csv

Interesting/Surprising Insights

- In the boxplot, the median satisfaction level for employees who leave the company is much lower than employees who do not.
- In the bar chart, we can see that people with very few or a lot of projects (2, 6, 7) tend to leave the company more than others.





Inferences

- Features that impact satisfaction level are number_projects, average_monthly_hours, time_spent_company and last_evaluation.
- Features that impact the attrition rate are satisfaction_level, time_spent_company, average_monthly_hours, last_evaluation

Challenges

In the classifiers, the indexed columns (feature engineered) turn out to be features with low importance.

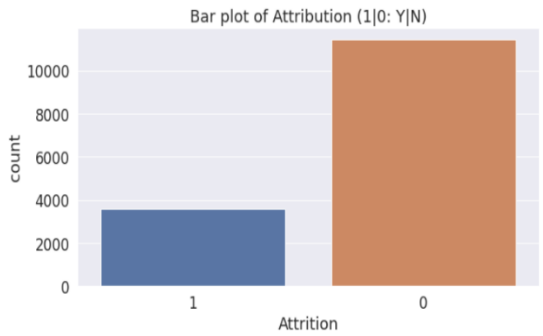
ROC and PR curves were not plotted for the support vector machine since there is no 'probability' column required to plot the two curves.

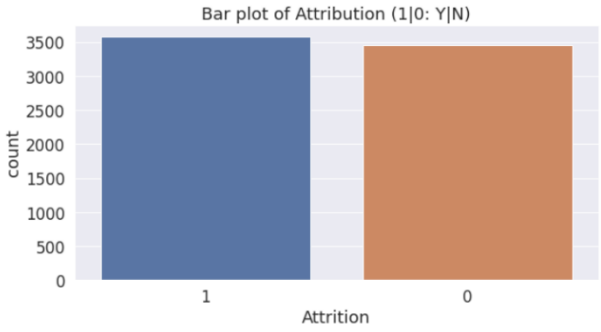
Data Collection and Cleaning

Dataset Pre-processing

Data cleaning was initiated by changing the datatype of some columns to their appropriate type. Added an additional column named “attrition” which is the same as the column “left”. This was created to be used later in classification models. Renamed ‘sales’ column to ‘department’. The data was found to be imbalanced hence, balancing was performed using weighted sampling. There are no missing values in the dataset.

```
root
 |-- satisfaction_level: float (nullable = true)
 |-- last_evaluation: float (nullable = true)
 |-- number_project: integer (nullable = true)
 |-- average_montly_hours: string (nullable = true)
 |-- time_spend_company: float (nullable = true)
 |-- Work_accident: string (nullable = true)
 |-- left: string (nullable = true)
 |-- promotion_last_5years: string (nullable = true)
 |-- department: string (nullable = true)
 |-- salary: string (nullable = true)
 |-- average_monthly_hours: float (nullable = true)
 |-- attrition: float (nullable = true)
```

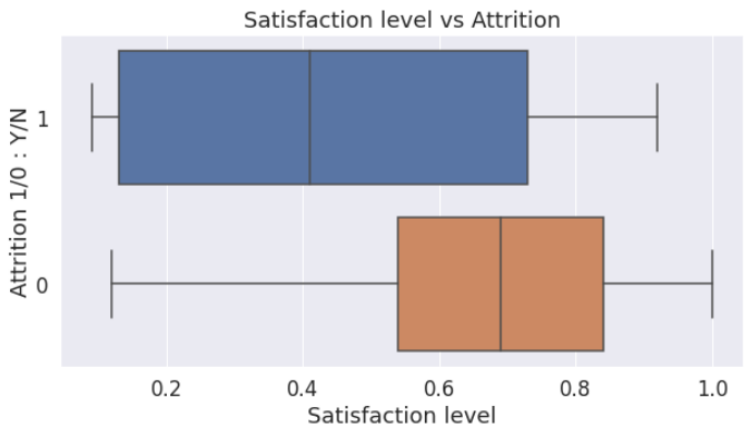




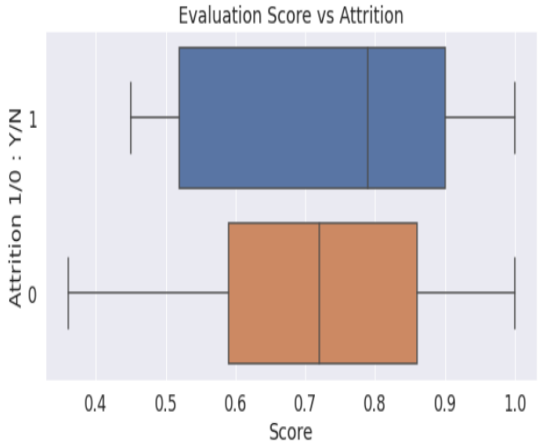
Data Exploration Insights

Data Insights

As seen in the figure, the median of the attrition=0 boxplot indicates that the employees who tend to not leave the company have a higher satisfaction level. This insight is backed by the classification models because ’satisfaction_level’ has a high feature importance.



Visualizations



Interesting facts you learnt from data

- From the bar plot above, it can be inferred that the employees assigned to 3/4 projects are less likely to leave the company than employees with a high or low number of projects.
- From the box plot above, the employees whose last evaluation score high tends to leave the company compared to people with lower evaluation scores indicating that better employees are more likely to leave the organization.

Statistical summary

As seen from the statistical table below, the means are widely distributed, this indicates the requirement to scale and hence has been done while creating the machine learning models.

summary	satisfaction_level	last_evaluation	number_project	time_spend_company	average_monthly_hours
count	14999	14999	14999	14999	14999
mean	0.6128335220371535	0.7161017401400102	3.80305353690246	3.498233215547703	201.0503366891126
stddev	0.24863065162689157	0.17116911257197398	1.2325923553183513	1.4601362305354808	49.94309937128406
min	0.09	0.36	2	2.0	96.0
max	1.0	1.0	7	10.0	310.0

Methodology

Method description

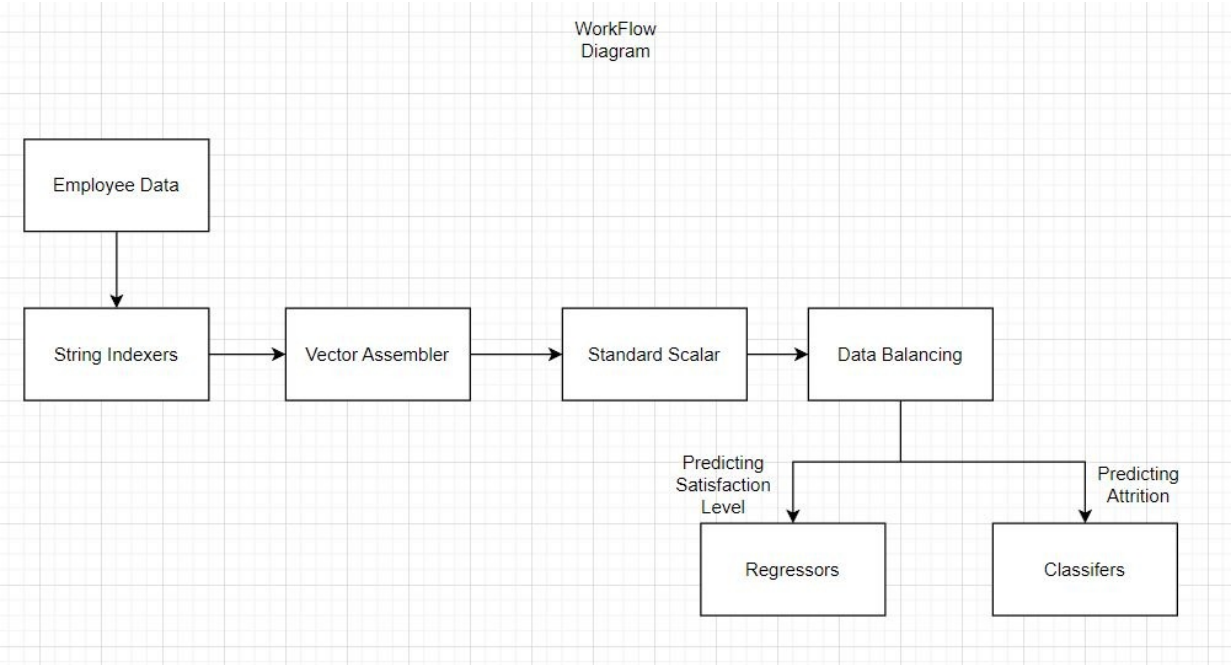
String data was indexed using string indexers. This indexed data was later one hot encoded. Vector assembler was used to create the feature column by combining all the predictors. These

features were later scaled. Output of the feature engineering pipe gives us a sparse vector of the scaled features.

Scoring metrics

- The business case requires the usage of both regressors and classifiers for predictive analytics. Different scoring metrics have been used to evaluate the models.
- Scoring metrics for regressors: Mean Squared Error (MSE), Root Mean Squared Error (RMSE)
- Scoring metrics for classifiers: Accuracy, Precision, Recall, F1 Score, Area Under ROC Curve

Workflow



Model Prediction

Regression Models

Linear Regression

The model type is a regressor. The predictor is 'satisfaction_level' and the scoring metric for this model is MSE and RMSE. Grid search was not performed for this model since the hyperparameters don't impact the model performance to a high degree.

MSE	RMSE
0.0605	0.2461.

Random Forest Regressor

The model type is a regressor. The predictor is 'satisfaction_level' and the scoring metric for this model is MSE and RMSE. Grid search was performed in this model and the hyperparameters were tuned to find the lowest MSE and RMSE.

MSE	RMSE
0.0253	0.1691

Gradient Boosting Tree Regressor

The model type is a regressor. The predictor is 'satisfaction_level' and the scoring metric for this model is MSE and RMSE. Grid search was performed in this model and the hyperparameters were tuned to find the lowest MSE and RMSE.

MSE	RMSE
0.0271	0.1390

The MSE and RMSE of all the regressors are low because the data to be predicted (satisfaction level) is in the range from 0 to 1.

Classification Models

Logistic Regression

The model type is a classifier. The predictor is 'attrition' and the scoring metric for this model are accuracy, precision, recall, F1 score and the area under ROC curve. Grid search was not performed for this model since the hyperparameters don't impact the model performance to a high degree.

Accuracy	Precision	Recall	F1 Score	Area under ROC
0.727022	0.716263	0.629179	0.669903	0.702519

Random Forest Classifier

The model type is a classifier. The predictor is 'attrition' and the scoring metric for this model are accuracy, precision, recall, F1 score and the area under ROC curve. Grid search was performed in this model and the hyperparameters were tuned to find the highest accuracy.

Accuracy	Precision	Recall	F1 Score	Area under ROC
0.976258	0.925411	0.978723	0.945668	0.946748

Gradient Boosting Tree Classifier

The model type is a classifier. The predictor is 'attrition' and the scoring metric for this model are accuracy, precision, recall, F1 score and the area under ROC curve. Grid search was performed in this model and the hyperparameters were tuned to find the highest accuracy.

Accuracy	Precision	Recall	F1 Score	Area under ROC
0.983464	0.937276	0.969416	0.954075	0.952285

Multilayer Perceptron Classifier

The model type is a classifier. The predictor is 'attrition' and the scoring metric for this model are accuracy, precision, recall, F1 score and the area under ROC curve. Grid search was performed in this model and the hyperparameters were tuned to find the highest accuracy.

Accuracy	Precision	Recall	F1 Score	Area under ROC
0.859220	0.933579	0.768997	0.84333	0.874107

Support Vector Machine

The model type is a classifier. The predictor is 'attrition' and the scoring metric for this model are accuracy, precision, recall, F1 score and the area under ROC curve. Grid search was performed in this model and the hyperparameters were tuned to find the highest accuracy.

Accuracy	Precision	Recall	F1 Score	Area under ROC
0.805747	0.812500	0.750760	0.780411	0.797497

Model Inference

Regression Models

The following regressors are used to predict the satisfaction_level of an employee.

Linear Regression

- The goal of the inference is to assess the impact of unique features on the predicting variable. Eventually, this information is used to tune the best model by removing the least important features from the predictor.
- Scores have been sorted based on absolute value to get the most important positive and negative features. More the positive score, more the positive impact and more the

negative score, more the negative impact. If the salary of an employee who earns less increases by 1 unit, then his/her satisfaction level will increase by 0.507289 units.

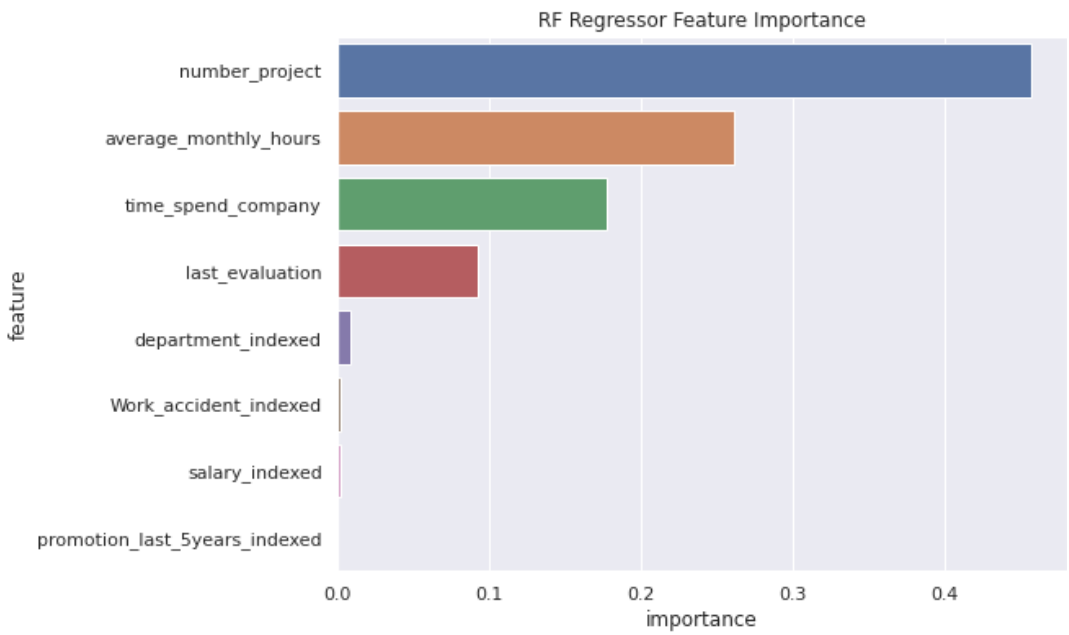
- Transformed the data using a standard scaler so regression coefficients representing real data could be compared between each other

name	score
en_salary_low	0.507289
en_department_RandD	-0.078224
en_department_accounting	-0.068710
en_salary_medium	-0.060745
en_department_marketing	-0.058500
en_department_hr	-0.052305
en_department_product_mng	0.050922

Random Forest Regressor

- The goal of the inference is to assess the impact of different features on the predicting variable by extracting the feature importance from the model. Eventually, this information is used to tune the best model by removing the least important features from the predictor
- Number of projects, average monthly hours, last evaluation and the time spent at the company are important to predict satisfaction level

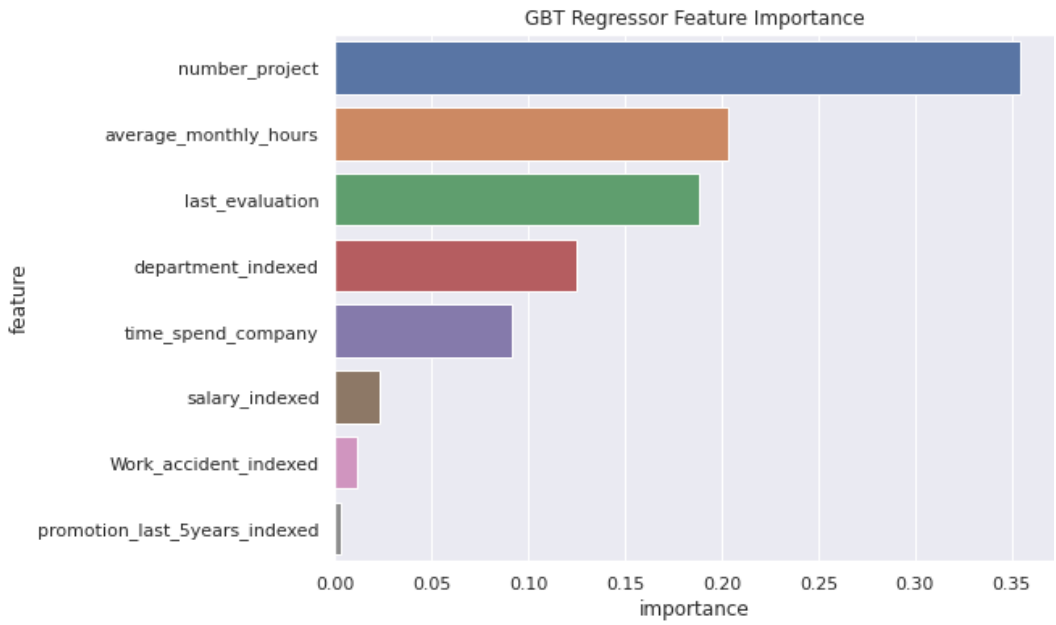
feature	importance
number_project	0.598368
average_monthly_hours	0.212040
last_evaluation	0.099163
time_spend_company	0.078560
department_indexed	0.006927
salary_indexed	0.002002
Work_accident_indexed	0.001666
promotion_last_5years_indexed	0.001274



Gradient Boosting Tree Regressor

- The goal of the inference is to assess the impact of different features on the predicting variable by extracting the feature importance from the model. Eventually, this information is used to tune the best model by removing the least important features from the predictor
- Number of projects, average monthly hours, last evaluation and the department are important to predict satisfaction level

feature	importance
number_project	0.354675
average_monthly_hours	0.203399
last_evaluation	0.188044
department_indexed	0.124558
time_spend_company	0.091646
salary_indexed	0.022922
Work_accident_indexed	0.011756
promotion_last_5years_indexed	0.003001



Classification Models

The following classifiers are used to predict the attrition of an employee.

Logistic Regression

- The goal of the inference is to find out the best hyperparameters for the model using grid search cv. Eventually, this information is used to tune the best model by removing the least important features from the predictor and hard coding the hyper parameter
- Number of projects, average monthly hours, last evaluation and the department are important to predict satisfaction level

name	score
en_promotion_last_5years_0	-4.368143
en_department_product_mng	2.158816
en_department_hr	1.984340
en_department_accounting	1.582516
en_department_marketing	1.365504
en_department_RandD	1.031911
number_project	0.922077

Random Forest Classifier

- The goal of the inference is to find out the best hyperparameters for the model using grid search cv. Eventually, this information is used to tune the best model by removing the least important features from the predictor and hard coding the hyper parameter
- Number of projects, average monthly hours, satisfaction level and the time spent at company are important to predict attrition

feature	importance
number_project	0.260984
time_spend_company	0.247045
satisfaction_level	0.227111
average_monthly_hours	0.133057
last_evaluation	0.115677
Work_accident_indexed	0.011285
salary_indexed	0.002957
department_indexed	0.001884
promotion_last_5years_indexed	0.000000

Gradient Boosting Tree Classifier

- The goal of the inference is to find out the best hyperparameters for the model using grid search cv. Eventually, this information is used to tune the best model by removing the least important features from the predictor and hard coding the hyper parameter
- Last Evaluation, average monthly hours, satisfaction level and the time spent at company are important to predict attrition

feature	importance
satisfaction_level	0.397553
time_spend_company	0.269411
last_evaluation	0.110133
average_monthly_hours	0.106771
department_indexed	0.052043
number_project	0.047483
salary_indexed	0.014291
Work_accident_indexed	0.002315
promotion_last_5years_indexed	0.000000

Multilayer Perceptron Classifier

- The goal of the inference is to find out the best hyperparameters for the model using grid search cv. Eventually, this information is used to tune the best model by removing the least important features from the predictor and hard coding the hyper parameter
- Since feature importance cannot be extracted for Multi-Layer Perceptron, feature importance from GBT classifier is used to determine the key features during the model tuning
- Last Evaluation, average monthly hours, satisfaction level and the time spent at company are important to predict attrition

Support Vector Machine

- The goal of the inference is to find out the best hyperparameters for the model using grid search cv. Eventually, this information is used to tune the best model by removing the least important features from the predictor and hard coding the hyper parameter
- Since feature importance cannot be extracted for Support Vector Machine, feature importance from GBT classifier is used to determine the important features during the model tuning
- Last Evaluation, average monthly hours, satisfaction le and the time spent at company are important to predict satisfaction level

Conclusion

Model Comparisions

Comparative Analytics – Regressors

The best regressor to predict satisfaction level is Random Forest as seen from the MSE below.

Model	Train MSE	Test MSE	Train RMSE	Test RMSE
Linear Regression	0.063075	0.060586	0.251148	0.246141
Random Forest Regressor	0.028603	0.025321	0.169124	0.159127
Gradient Boosting Tree Regressor	0.139083	0.027179	0.139083	0.164859

Comparative Analytics – Classifiers

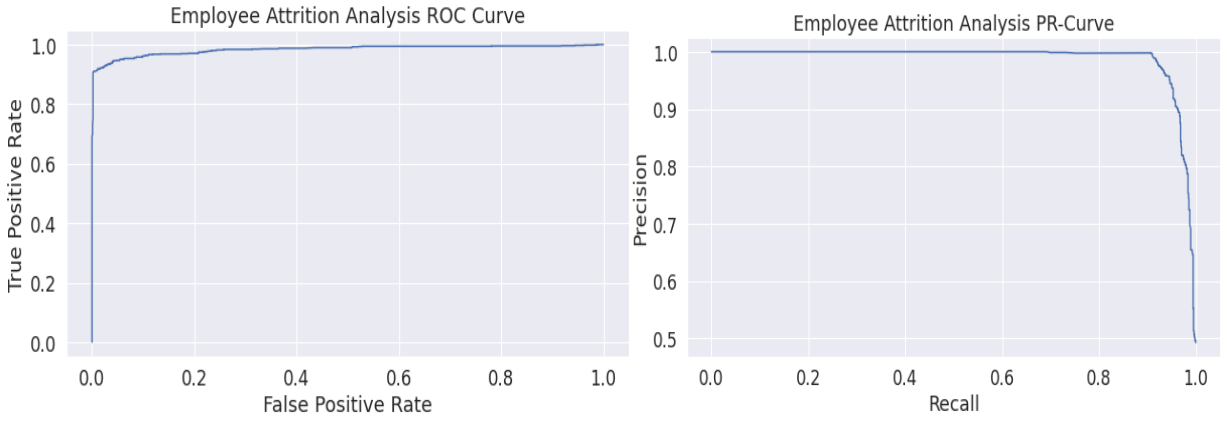
The best classifier to predict attrition is the Gradient Boosting Tree as seen from the scoring metrics below.

Model	Train Accuracy	Test Accuracy	Train Recall	Test Recall	Train Precision	Test Precision	Train F1 Score	Test F1 Score	AUC
Logistic Regression	0.822090	0.832309	0.723716	0.775076	0.785980	0.777439	0.753564	0.776256	0.783904
Random Forest Classifier	0.978565	0.976258	0.981418	0.978723	0.925311	0.914773	0.952539	0.945668	0.946748
Gradient Boosting Tree Classifier	0.996332	0.983464	0.987286	0.969416	0.964183	0.937276	0.975598	0.953075	0.952285
SVM Classifier	0.793034	0.805747	0.711002	0.750760	0.806434	0.812500	0.755717	0.780411	0.797497
Multilayer Perceptron Classifier	0.860393	0.859220	0.755012	0.768997	0.947821	0.933579	0.840501	0.843333	0.874107

Area Under ROC Curves

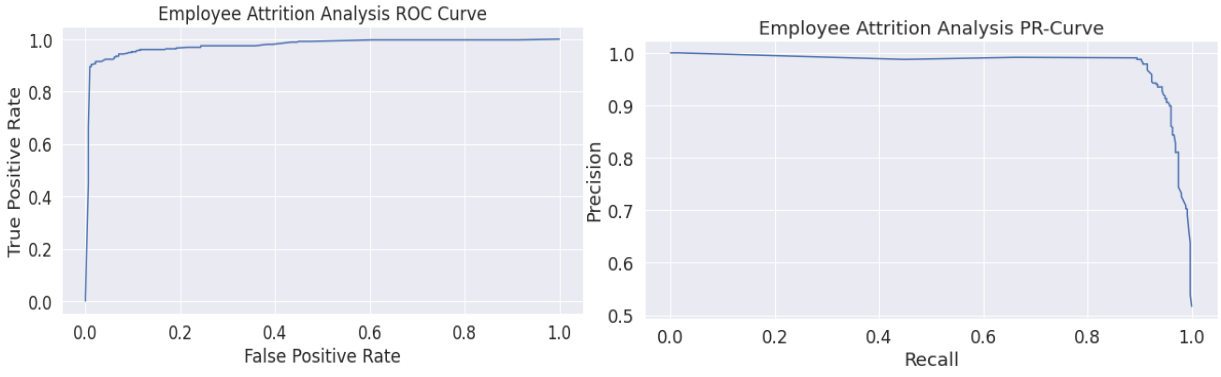
GBT Classifier:

The AUC for the GBT classifier is 0.952



Random Forest Classifier:

The AUC for the GBT classifier is 0.947



Results

From the exploratory data analytics, there was a hunch that satisfaction level would be an important predictor to predict attrition. From the feature importance of the classifiers, this hunch was confirmed, and satisfaction level turned out to be the most important predictor. Another finding was that the number of projects also had a high impact to predict attrition.

For this dataset, gradient boosting tree classifier is the best machine learning model to predict attrition and random forest regressor is the best model to predict satisfaction level.

Appendix

1. https://thecleverprogrammer.com/2020/07/12/employee-turnover-prediction/
2. https://raw.githubusercontent.com/rnvira/Employee-Attrition/main/Employee_Attrition.csv
3. https://spark.apache.org/docs/latest/ml-pipeline.html
4. https://spark.apache.org/docs/3.0.1/api/python/pyspark.ml.html