VICKNESWARAN KEERTHTHANAN

18.09.2023

Employee Attrition Classification Project Report

Introduction

This report presents the findings and analysis of a classification project aimed at predicting employee attrition. The dataset used for this analysis contains information on various attributes of employees, including demographics, job-related factors, and performance metrics.

Methodology

Step 1: Problem Statement and Approach

The goal of this project was to perform a classification task to predict employee attrition. This was achieved by a machine learning model.

Step 2: Data Preprocessing and Feature Selection

Data preprocessing is a crucial step in preparing the dataset for analysis. It involves several sub-steps

2.1 Handling Missing Values

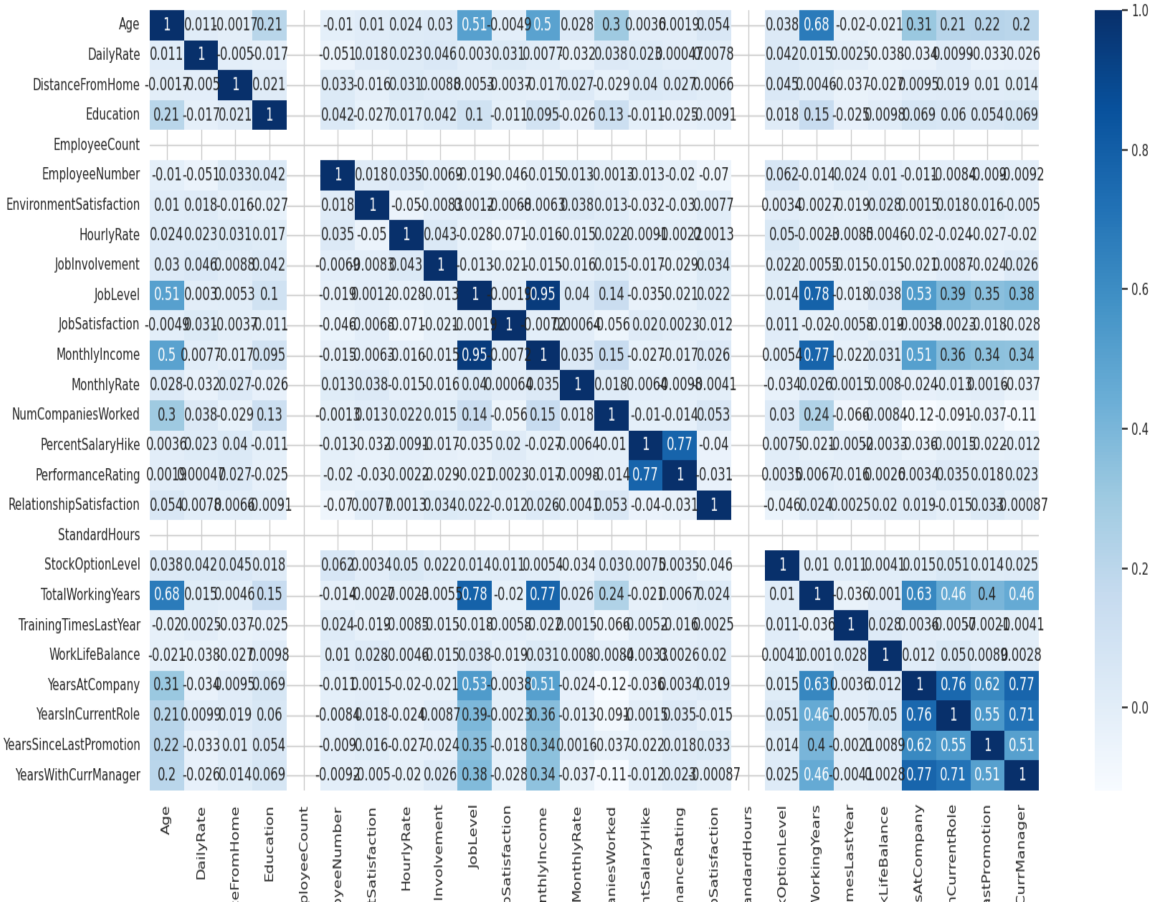
The dataset was inspected for any missing or incomplete data points. No missing or incomplete data points were identified during the inspection of the dataset. Therefore, no imputation or handling of missing values was required, ensuring that the dataset remained complete

2.2 Feature Selection

2.2.1 Pearson Correlation (Numeric Features)

Feature selection was performed to identify and address potential multicollinearity among numeric features. This process involved the use of Pearson correlation solely among the numeric features themselves. Features exhibiting strong linear correlations with each other were identified and subsequently removed

[MonthlyIncome, TotalWorkingYears, PercentSalaryHike] These features were removed with help of below matrix

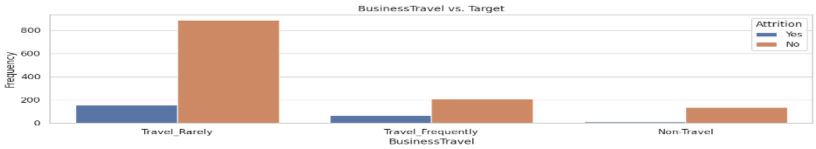


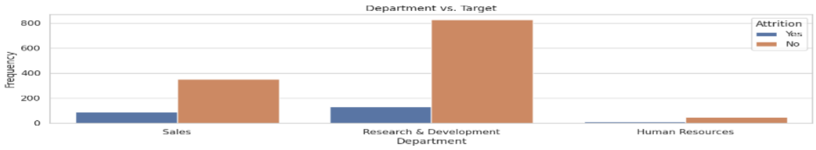
2.2.2 Visualization and Chi-Square Test (Categorical Features)

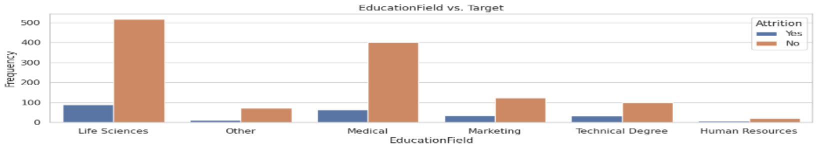
Categorical variables were evaluated using visualization techniques to explore their relationship with attrition. Additionally, the Chi-square test of independence was applied to determine the association between categorical features and attrition. Features exhibiting significant associations were selected for inclusion in the final model.

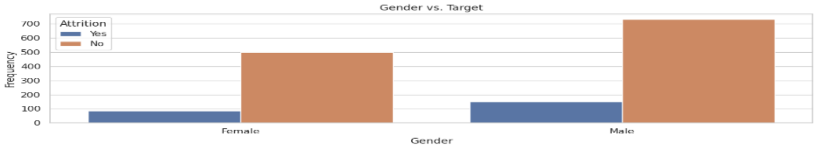
'Gender', 'Over18' these features were removed by using below charts and results of hypothesis test

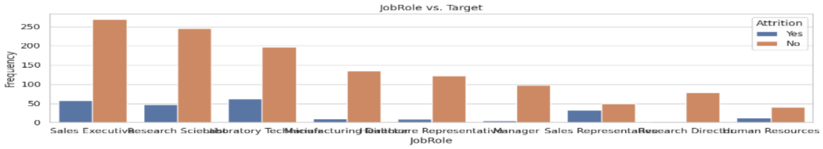
The feature selection process aimed to reduce multicollinearity, which occurs when two or more features are highly correlated with each other, potentially leading to unstable model estimates. Removing such features can improve the stability and interpretability of the final model.

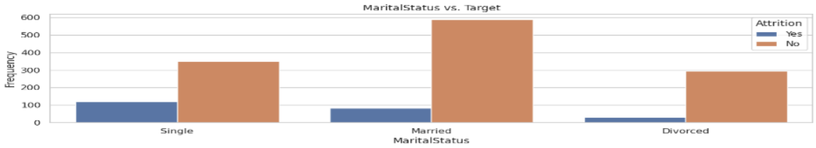

BusinessTravel vs. Target
Frequency
0
200
400
600
800
Travel_Rarely
Travel_Frequently
Non-Travel
BusinessTravel
Attrition
Yes
No

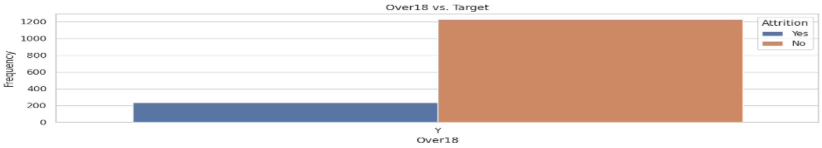

Department vs. Target
Frequency
0
200
400
600
800
Sales
Research & Development
Human Resources
Department
Attrition
Yes
No

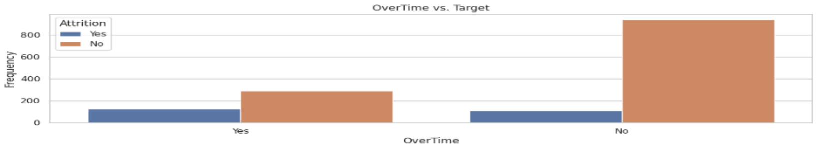

EducationField vs. Target
Frequency
0
100
200
300
400
500
Life Sciences
Other
Medical
Marketing
Technical Degree
Human Resources
EducationField
Attrition
Yes
No


Gender vs. Target
Frequency
0
100
200
300
400
500
600
700
Female
Male
Gender
Attrition
Yes
No


JobRole vs. Target
Frequency
0
50
100
150
200
250
Sales Executive
Research Scientist
Laboratory Technician
Manufacturing Director
Healthcare Representative
Manager
Sales Representative
Research Director
Human Resources
JobRole
Attrition
Yes
No


MaritalStatus vs. Target
Frequency
0
100
200
300
400
500
600
Single
Married
Divorced
MaritalStatus
Attrition
Yes
No


Over18 vs. Target
Frequency
0
200
400
600
800
1000
1200
Y
Over18
Attrition
Yes
No


OverTime vs. Target
Frequency
0
200
400
600
800
Yes
No
OverTime
Attrition
Yes
No

2.3 Data Splitting

To effectively evaluate the performance of the machine learning model, the dataset was divided into two subsets:

2.3.1 Training Set (80%)

The larger portion, constituting 80% of the dataset, was allocated for training the model. This subset was used to teach the model to recognize patterns in the data.

2.3.2 Testing Set (20%)

The remaining 20% of the dataset was reserved for testing the model. This subset was used to assess how well the model generalized to new, unseen data.

This division ensured a robust evaluation of the model's performance on data it had not been exposed to during training.

2.4 Encoding the Categorical column

Categorical variables including 'BusinessTravel', 'Department', 'EducationField', 'JobRole', 'MaritalStatus', and 'OverTime' were encoded using one-hot encoding. This technique ensures that categorical variables are appropriately represented for the machine learning model. The ColumnTransformer was used to apply this transformation selectively to categorical features, while leaving numeric features untouched.

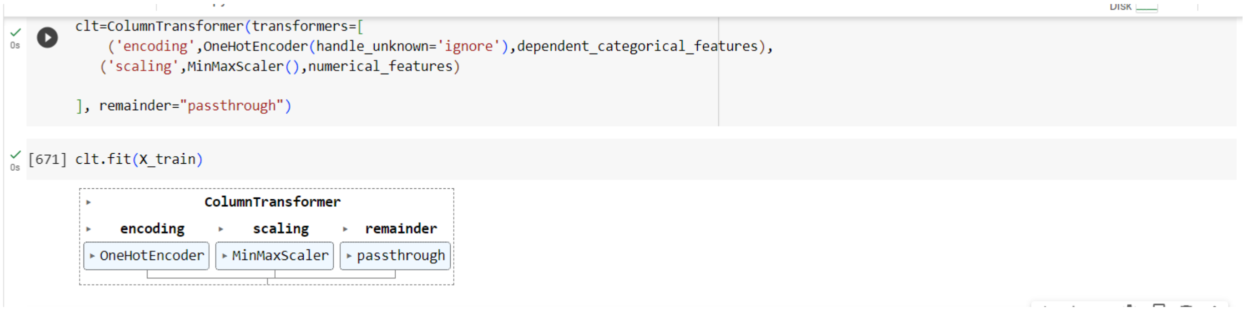
This step facilitated the integration of categorical variables into the model training process, allowing for a comprehensive analysis of their impact on attrition prediction

2.5 Scaling Numerical Features

Numerical features, such as age or monthly income, were standardized using the Min-Max scaling technique.

This process ensured that all numeric features were on a similar scale, preventing any single feature from dominating the model training process.

This step further optimized the data for model training, enhancing the performance and stability of the final predictive model.



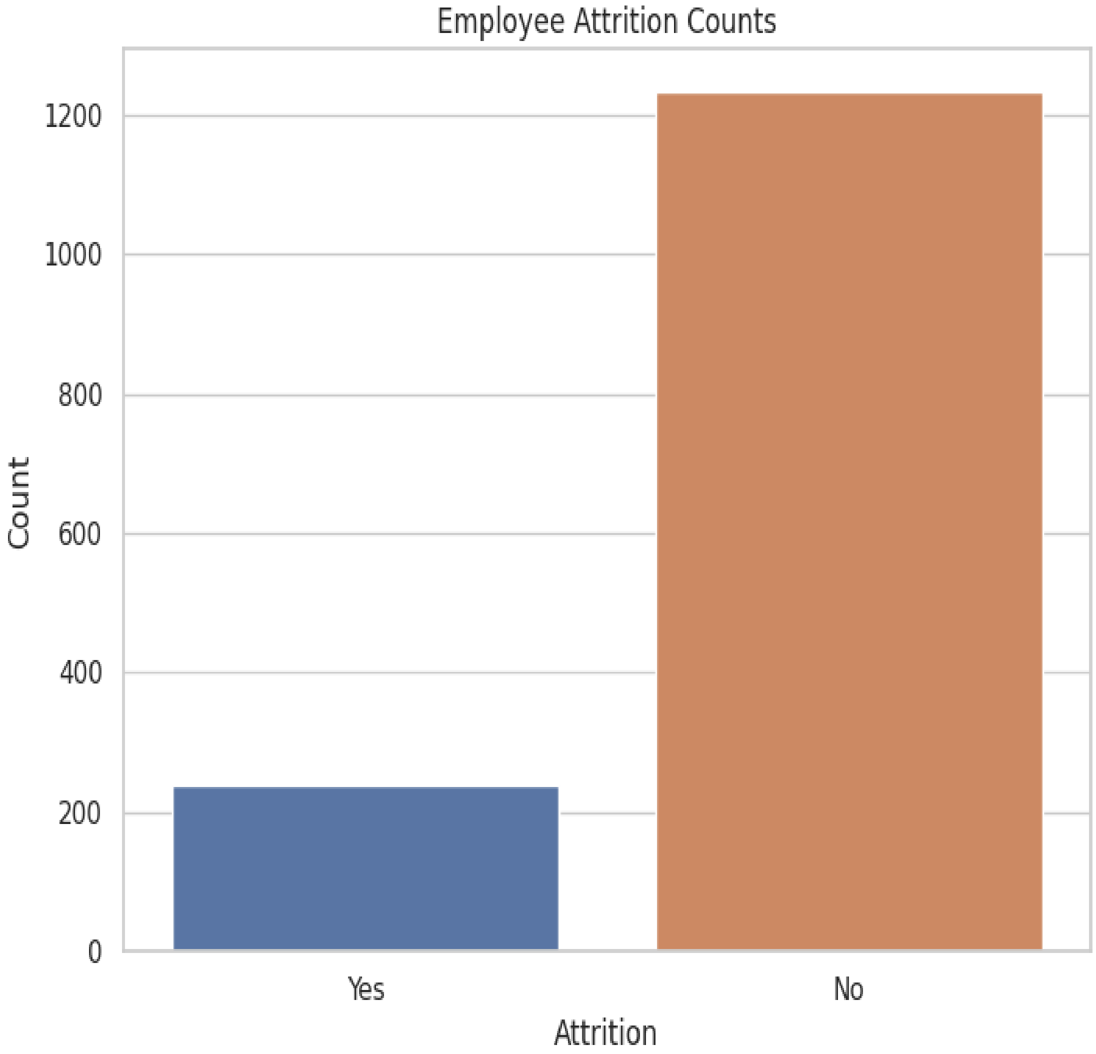
```
clt=ColumnTransformer(transformers=[
    ('encoding',OneHotEncoder(handle_unknown='ignore'),dependent_categorical_features),
   ('scaling',MinMaxScaler(),numerical_features)

], remainder="passthrough")
```

```
[671] clt.fit(X_train)
```

2.6 Addressing Class Imbalance

Upon encoding the target(Attrition) variable, it was observed that the dataset exhibited class imbalance. To mitigate this issue, undersampling was used. This technique involves reducing the size of the majority class to create a more balanced distribution between the classes. By doing so, aimed to ensure that the model's training process was not biased towards the majority class, thus enhancing its ability to generalize effectively to both classes.

The application of undersampling contributed to a more balanced training set, allowing the model to learn from both classes with equal importance.



Step 3: Model Training and Selection

After completing the data preprocessing and feature selection steps, the next phase involved training and selecting the most appropriate machine learning model for predicting employee attrition.

To determine the optimal model, several algorithms were considered:

Random Forest Classifier

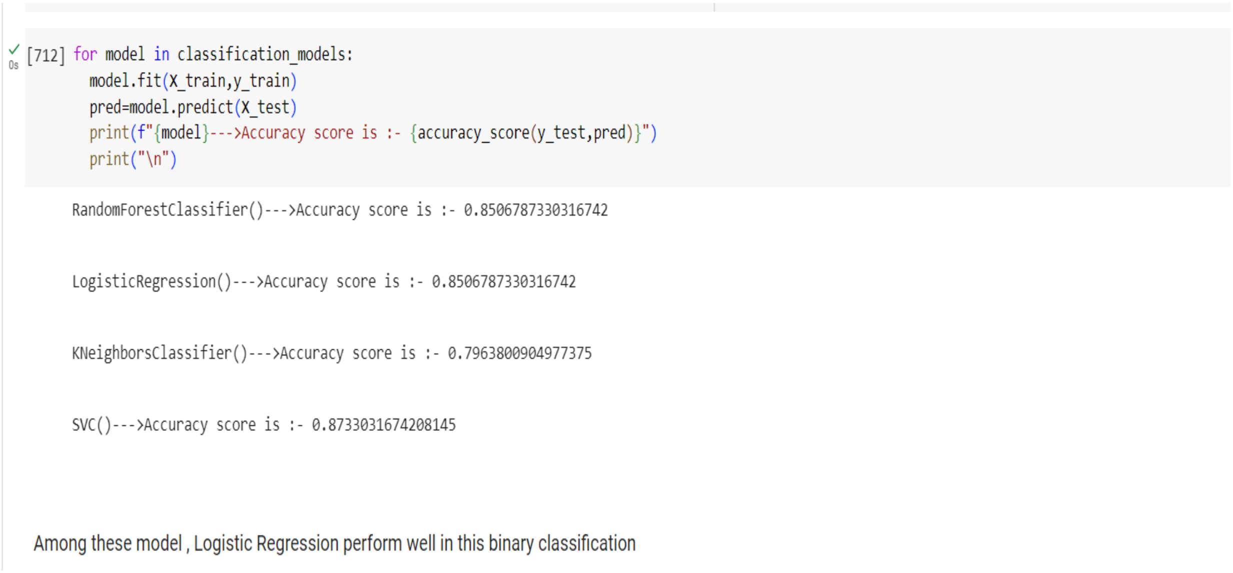
Support Vector Classifier (SVC)

Logistic Regression

K-Nearest Neighbors (KNN)

Each model was initially trained with default parameters to establish a baseline performance.

It was observed that the Logistic Regression model outperformed the other models in terms of its predictive accuracy and overall performance on the given dataset. The decision to proceed with Logistic Regression was based on its ability to effectively capture the underlying relationships in the data, particularly in the context of predicting employee attrition. Additionally, Logistic Regression provided a balance between simplicity and predictive power, making it a suitable choice for this specific task.



```
[712] for model in classification_models:
        model.fit(X_train,y_train)
        pred=model.predict(X_test)
        print(f"{model}--->Accuracy score is :- {accuracy_score(y_test,pred)}")
        print("\n")

    RandomForestClassifier()--->Accuracy score is :- 0.8506787330316742


    LogisticRegression()--->Accuracy score is :- 0.8506787330316742


    KNeighborsClassifier()--->Accuracy score is :- 0.7963800904977375


    SVC()--->Accuracy score is :- 0.8733031674208145
```

Among these model , Logistic Regression perform well in this binary classification

Step 4: Hyperparameter Tuning

Given the selection of Logistic Regression as the chosen model, the next phase involved fine-tuning its hyperparameters. This step is crucial for optimizing the model's performance and achieving the best possible predictive accuracy.

4.1 Initial Hyperparameters

An initial set of hyperparameters was chosen for the Logistic Regression model, including parameters related to regularization , penalty , and solver options.

4.2 Grid Search and Cross-Validation

Grid search, a technique for systematically testing multiple combinations of hyperparameters, was employed. This involved defining a grid of hyperparameter values and exhaustively evaluating the model performance for each combination.

Cross-validation, specifically k-fold cross-validation, was utilized to robustly assess model performance across different subsets of the training data

4.3 Outcome

Despite efforts, no significant accuracy improvement was achieved.

4.4 Final Model Selection

Retained Logistic Regression with default settings.

Step 5: Model Evaluation

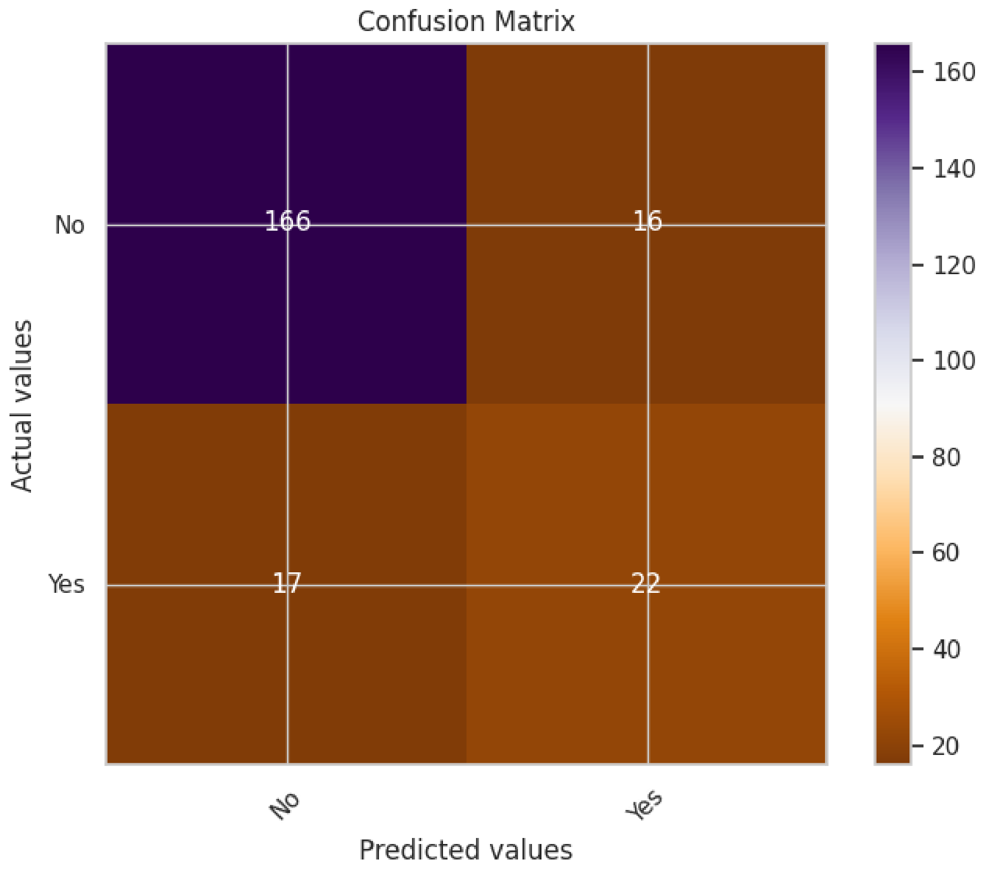
5.1 Accuracy Score

The model achieved an accuracy score of 86%, indicating the proportion of correctly classified instances in the test set.

5.2 Confusion Matrix

A confusion matrix was utilized to provide a detailed view of the model's performance. It breaks down predictions into true positives, true negatives, false positives, and false negatives, enabling a comprehensive assessment of the model's predictive capabilities.

This step allowed for a thorough understanding of how well the model performed in distinguishing between employees who stayed and those who left the organization. The combination of accuracy score and confusion matrix provided valuable insights into the model's strengths and areas for potential improvement.



Visualizations (Power BI)

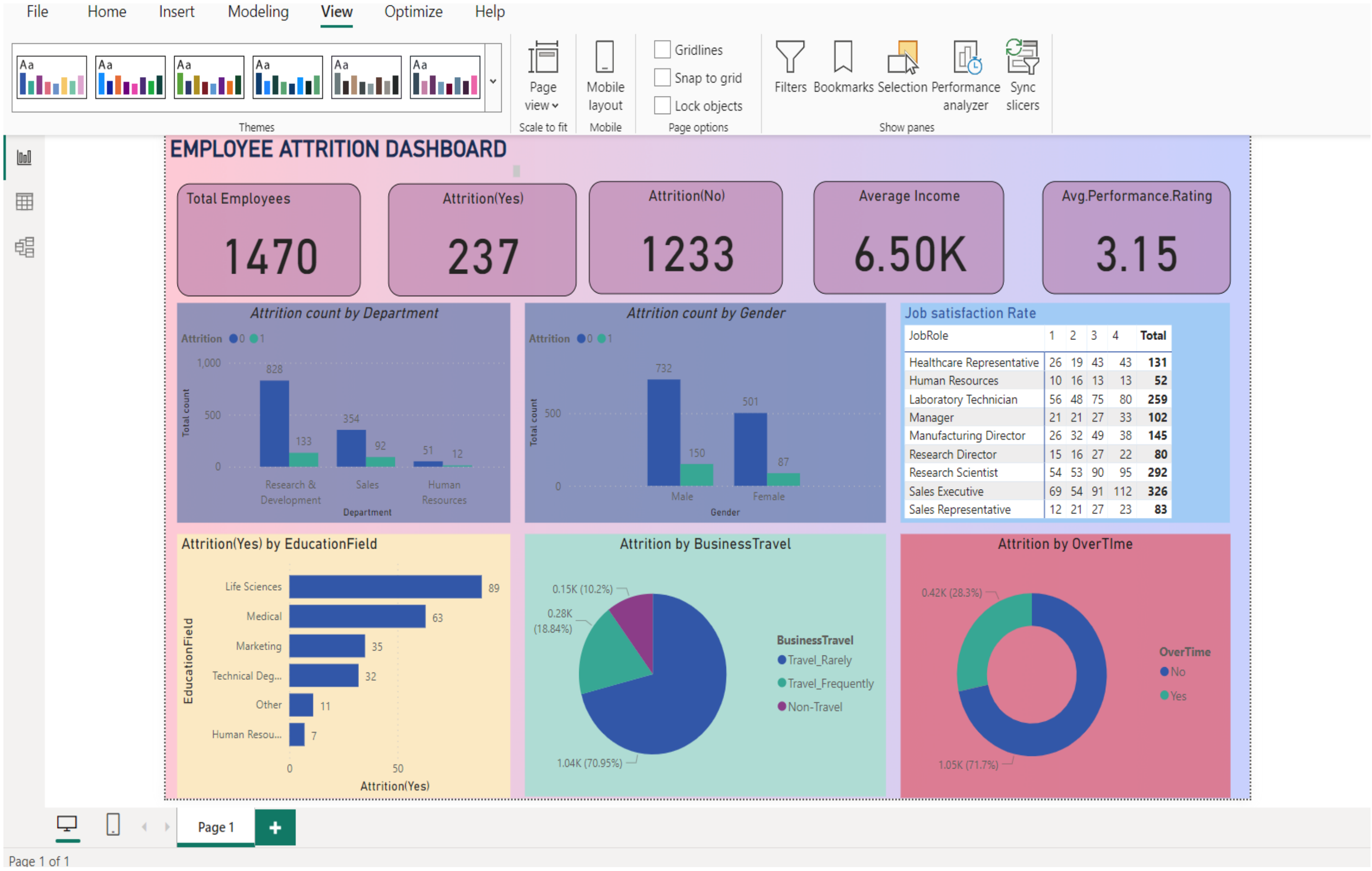
Attached is a screenshot of the Power BI dashboard summarizing key insights derived from the employee attrition dataset. The visualizations provide a clear overview of trends, patterns, and important metrics related to attrition. Some of the highlights include:

Demographic Analysis: Visualizations depicting attrition breakdowns by factors such as department, education field, job role, and marital status, providing insights into potential areas of focus for retention efforts.

Business Travel Impact: Visualization illustrating the impact of business travel on attrition, providing valuable insights into how travel may influence employee attrition

Over Time Impact: A visual representation of the effect of overtime work on attrition, shedding light on potential correlations between extended work hours and employee retention.

The Power BI dashboard serves as a dynamic tool for exploring and analyzing the data, allowing for a deeper understanding of the factors contributing to employee attrition within the organization.



EMPLOYEE ATTRITION DASHBOARD

Total Employees 1470

Attrition(Yes) 237

Attrition(No) 1233

Average Income 6.50K

Avg.Performance.Rating 3.15

Attrition count by Department

Attrition count by Gender

Job satisfaction Rate

JobRole	1	2	3	4	Total
Healthcare Representative	26	19	43	43	**131**
Human Resources	10	16	13	13	**52**
Laboratory Technician	56	48	75	80	**259**
Manager	21	21	27	33	**102**
Manufacturing Director	26	32	49	38	**145**
Research Director	15	16	27	22	**80**
Research Scientist	54	53	90	95	**292**
Sales Executive	69	54	91	112	**326**
Sales Representative	12	21	27	23	**83**

Attrition(Yes) by EducationField

Attrition by BusinessTravel

Attrition by OverTIme