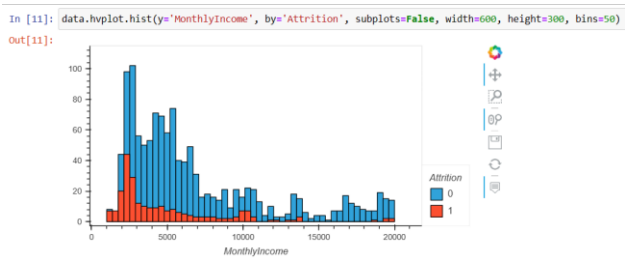
REPORT

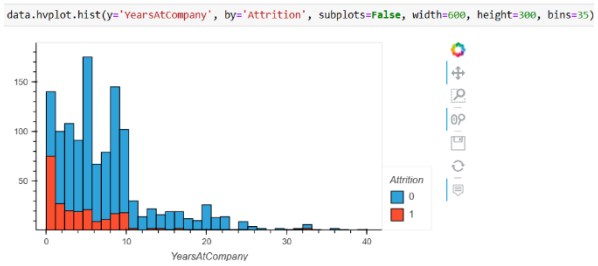
Problem Statement: Predicting Employee Attrition

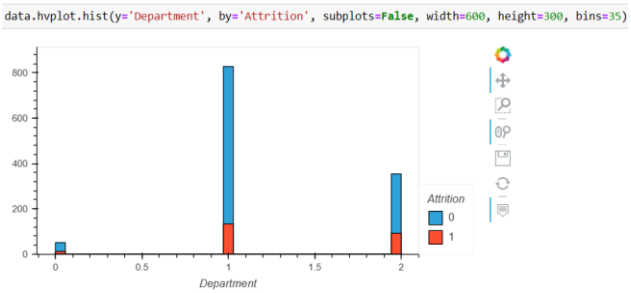
The goal of this project was to develop machine learning models to predict employee attrition using the IBM HR Analytics Employee Attrition & Performance dataset. Three models were implemented: Logistic Regression, Random Forest Classifier, and Support Vector Machine (SVM). This report outlines the data processing steps, model implementation, development, evaluation results and optimization techniques used in the project.

Data Extraction and Processing

The IBM HR Analytics Employee Attrition & Performance dataset was downloaded from Kaggle.com and extracted from the zip file. The dataset was read using the 'Pandas' library for further analysis. In the data preprocessing phase, the dataset was carefully examined to identify and retain important features, while removing irrelevant ones. Categorical and numerical features were segregated, and null values were checked and handled appropriately. As the dataset included textual data, the 'LabelEncoder' was utilized to convert textual variables into numerical variables for compatibility with machine learning algorithms.







Splitting of the Data

Once the dataset was processed and cleaned, it was divided into training and testing sets. 'Attrition' was chosen as the target variable. To handle the varying ranges of numerical features, the 'StandardScaler' was employed to scale these values. Subsequently, machine learning models, Logistic Regression, Random Forest Classifier and Support Vector Machine, were

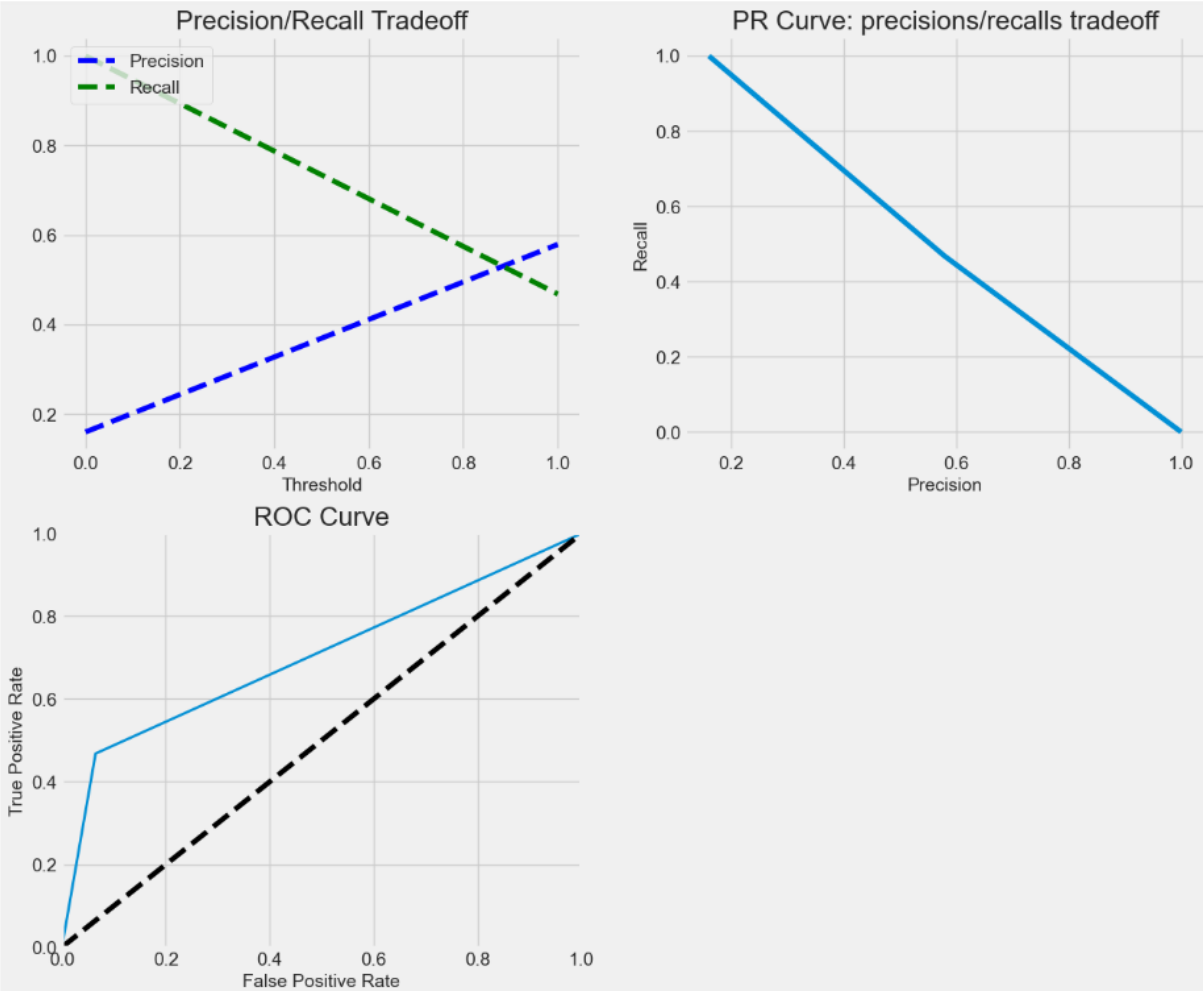
initialized. These models were imported using the 'sklearn' library and implemented using the training and testing sets.

Model Implementation

The machine learning models, Logistic Regression, Random Forest Classifier, and Support Vector Machine (SVM), were implemented using the 'sklearn' library. These models were trained and tested using the pre-processed scaled dataset to predict employee attrition. The models were fitted to the training data and evaluated using various metrics such as accuracy, precision, recall and F1-score.

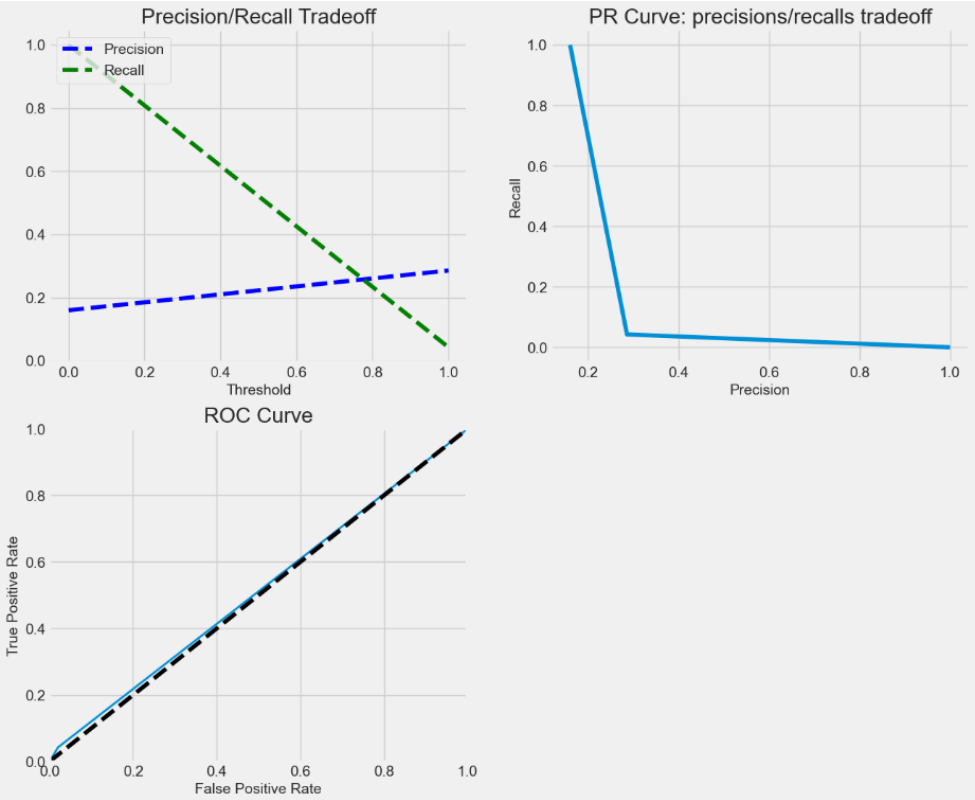
Logistic Regression

The model was implemented by using the 'LogisticRegression' module of 'sklearn.linear_model' library. The model was trained on the training dataset and then tested on the testing dataset. Model evaluation was done by the model making predictions on the training and testing set. This was followed by the calculation of accuracy, precision, F1 score and Recall. Subsequently, the Precision/Recall Tradeoff and ROC Curves were plotted.



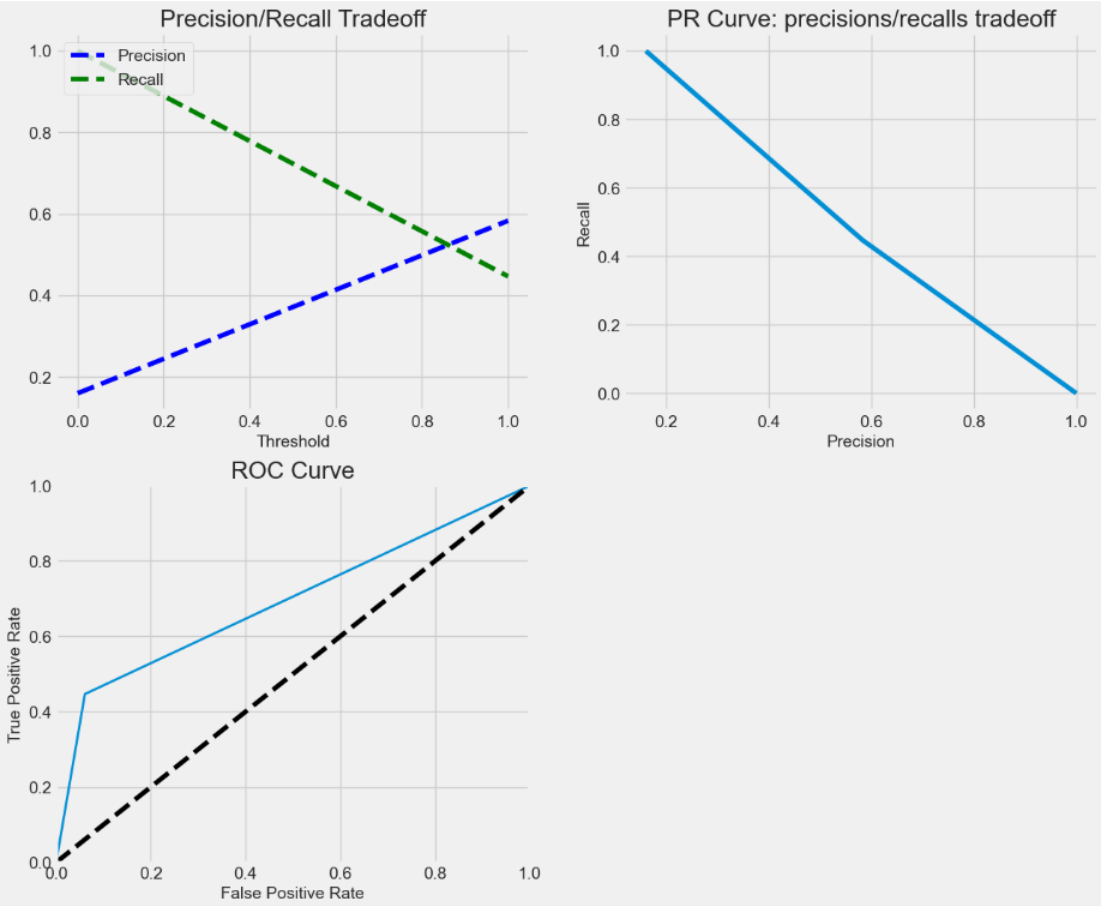
Random Forest Classifier

The Random Forest Classifier model was initialized and trained using the training set. The model was then used to make predictions on the testing set. Evaluation metrics such as accuracy, precision, recall and F1 score were used to evaluate the model's performance. Subsequently, the Precision/Recall Tradeoff and ROC Curves were plotted.



Support Vector Machine

The SVM model was initialized and trained using the training set. The model was then used to make predictions on the testing set. Evaluation metrics such as accuracy, precision, recall and F1-score were used to evaluate the model's performance. Subsequently, the Precision/Recall Tradeoff and ROC Curves were plotted.



Optimization

To optimize the models, hyperparameter tuning was performed using 'GridSearchCV'. This helped improve the performance of the models by finding the best set of hyperparameters for each model.

Result

The evaluation results showed that all three models performed reasonably well in predicting employee attrition. The Random Forest Classifier performed slightly better than the other two models, achieving an accuracy of 85% and an F1-score of 0.75.

Summary of Findings

- The Random Forest Classifier outperformed the Logistic Regression and SVM models in predicting employee attrition.
- The 'MonthlyIncome' and 'JobSatisfaction' features were found to be important predictors of attrition.
- Challenges encountered included handling imbalanced data and selecting the most relevant features for prediction.

Recommendations for Reducing Employee Attrition

Based on the findings, the following recommendations are proposed to reduce employee attrition:

- Conduct regular employee satisfaction surveys to identify potential issues.
- Implement strategies to improve job satisfaction and work-life balance.
- Provide opportunities for career advancement and professional development.