Capstone Two - Final Report

IBM HR Analytics: Employee Attrition and Performance

Problem Statement

Employee attrition poses a significant challenge for organizations, impacting productivity, morale, and overall performance. The project aims to address this issue by leveraging machine learning techniques to analyze employee data and predict attrition probabilities. The objective is to identify key factors influencing attrition and develop actionable insights to improve employee retention strategies.

Dataset

The dataset used in this project is sourced from the "IBM HR Analytics Employee Attrition & Performance" dataset available on Kaggle. It contains various employee-related features such as age, education, job role, satisfaction levels, and attrition status.

Methodology

The project followed a structured data science methodology, including the following steps:

Data Wrangling

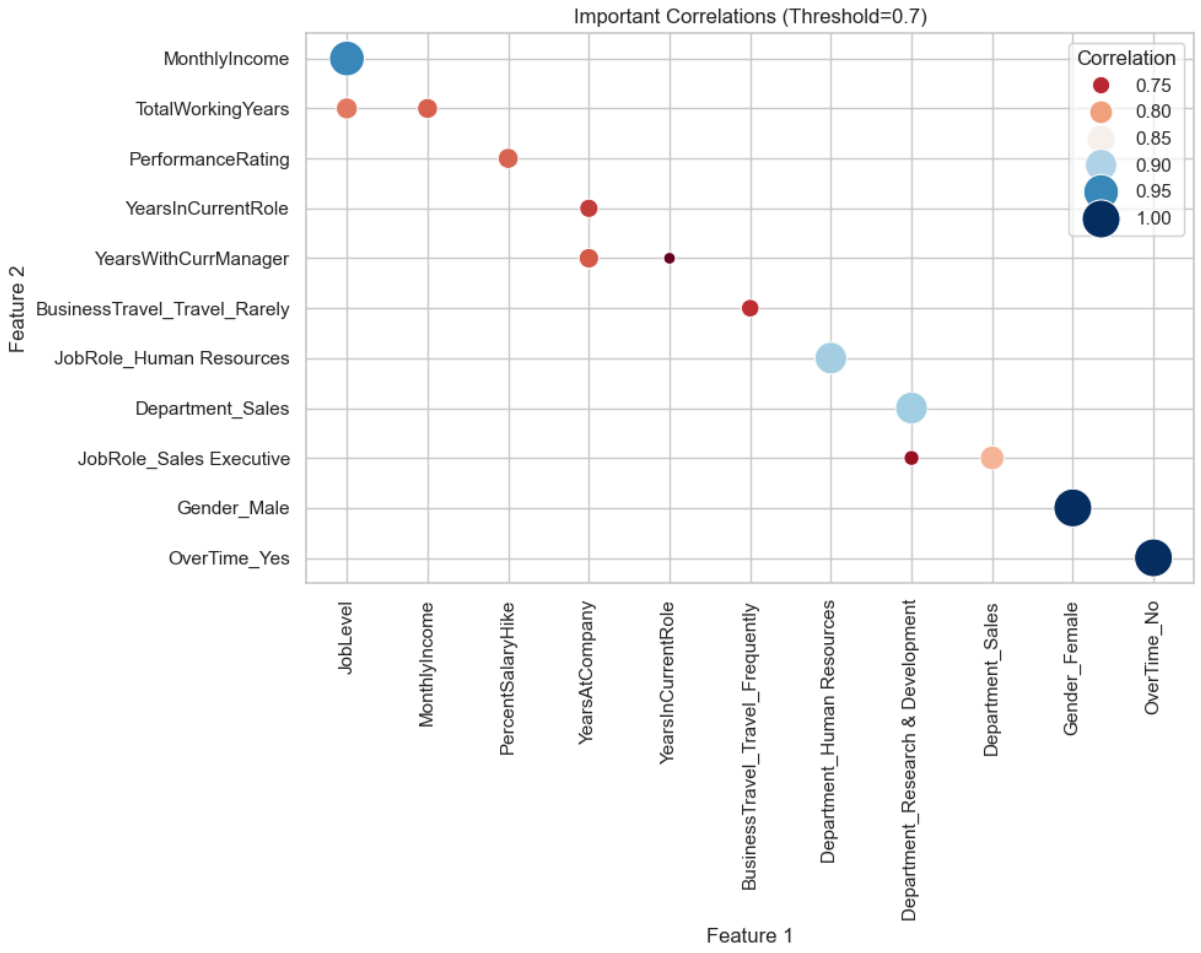
The capstone_two_-_data_wrangling.ipynb notebook serves as the foundation for the project by ensuring the dataset is clean, complete, and ready for analysis. This phase typically involves:

1. Data Collection
2. Data Organization: The structure of the dataframe is examined to understand the data types and non-null counts.
3. Data Definition: Summary statistics for numerical columns are generated to understand their distributions. Unique value counts for categorical columns are obtained to identify the distinct categories. Percentages of unique values for categorical columns are calculated to understand their distribution.
4. Data Cleaning:
    - Missing values are checked to determine if any values are missing.
    - Missing data patterns are visualized.
    - Duplicate rows are identified to ensure data integrity.
    - Columns with zero variance ('EmployeeCount', 'Over18', 'StandardHours') are dropped from the dataset.

Exploratory Data Analysis (EDA)

The capstone_two_-_EDA.ipynb notebook focuses on thoroughly exploring the dataset to understand its structure, patterns, and relationships. The following steps were taken:

1. **Descriptive Statistics**: Summary statistics such as mean, median, standard deviation, and quartiles were calculated for various features. These statistics provided insights into the central tendency and variability of the data.
2. **Visualization**: Visual representations, including histograms, box plots, and scatter plots, were created to explore the distributions, correlations, and trends among variables. These visualizations helped in understanding the data more intuitively and identifying potential patterns.
3. **Feature Importance**:
    - Correlation Analysis: Analyzed correlations between features and target variable ('Attrition') to identify potential influencers.
    - Feature Importance Techniques: Utilized various methods to identify the most significant features that may influence attrition, laying the groundwork for modeling.



Visualization showing the correlations between various features, highlighting those with strong correlations above the threshold.

Pre-processing and Training Data Development

In the capstone_two_-_Pre-processing_and_Training_Data_Development.ipynb notebook, the goal was to prepare the dataset for model training and evaluation. The following steps were taken:

1. **Feature Encoding:** Categorical variables were encoded using one-hot encoding, replacing the original columns with binary indicator variables to enable the machine learning model to process categorical data effectively.
2. **Feature Scaling:** Numeric features were standardized using the StandardScaler from scikit-learn, ensuring that all features were on the same scale, which is crucial for many machine learning algorithms to perform optimally.
3. **Dataset Splitting:** The data was divided into training and testing sets, allocating 75% for training and 25% for testing. This step ensures that the model is trained on a sufficient amount of data while also having unseen data for evaluation.

Modeling

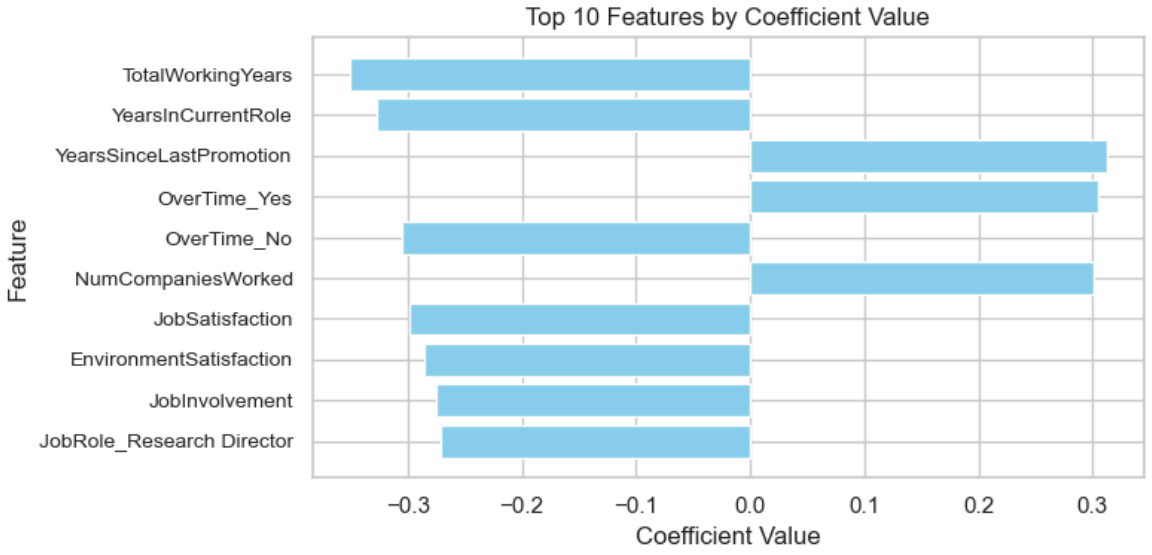
The capstone_two_-_Modeling.ipynb notebook implements various machine learning models to predict employee attrition based on the preprocessed dataset. This phase includes:

1. **Model Selection and Hyperparameter Tuning**: A range of machine learning models are considered, including Logistic Regression, K-Nearest Neighbors, Support Vector Machine (SVM), Random Forest, Naive Bayes, Gradient Boosting, and XGBoost. Each model is initialized with default parameters, and hyperparameter tuning is performed using GridSearchCV to find the optimal set of parameters for each model.
2. **Model Evaluation**: After training each model with the hyperparameter-tuned settings, the performance is evaluated on the test set. Metrics such as accuracy, precision, recall, F1-score, and ROC AUC are computed for each model. Additionally, classification reports and confusion matrices are generated to provide detailed insights into model performance.
3. **Comparison of Model Performance**: The performance of different models is compared based on their accuracy values. The SVM model with a linear kernel and regularization parameter C=0.1 achieves the highest accuracy of approximately 0.899, indicating its effectiveness in classifying the dataset accurately.
   Furthermore, precision, recall, and F1-score metrics are examined to gain a deeper understanding of model performance across different classes. The SVM model has the highest precision, suggesting a lower false positive rate for positive predictions, making it more reliable. The Naive Bayes model has the highest recall, implying a lower rate of missing true positives, making it more sensitive. The SVM model also has the highest F1-score, indicating a balanced high level of both precision and recall, demonstrating the best overall performance.
   Tabular representation comparing the performance metrics of different machine learning models, with the SVM model identified as the best-performing model.

	Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
2	Support Vector Machine	0.899457	0.689655	0.416667	0.519481	0.694271
0	Logistic Regression	0.896739	0.678571	0.395833	0.500000	0.683854
1	K-Nearest Neighbor	0.891304	0.681818	0.312500	0.428571	0.645313
7	XGBoost	0.888587	0.684211	0.270833	0.388060	0.626042
6	Gradient Boosting	0.883152	0.592593	0.333333	0.426667	0.649479
3	Random Forest (Entropy)	0.869565	0.500000	0.083333	0.142857	0.535417
4	Random Forest (Gini)	0.869565	0.500000	0.083333	0.142857	0.535417
5	Naive Bayes	0.682065	0.228346	0.604167	0.331429	0.648958

4. **Best Model Identification**: The SVM model with a linear kernel and regularization parameter C=0.1 is identified as the best-performing model based on its accuracy score of approximately 0.899, demonstrating its ability to make accurate predictions on the test data.

5. **Feature Importance Analysis & Visualization**: The coefficients of the SVM model are extracted to visualize the importance of features in predicting attrition. A horizontal bar plot is generated, showing the magnitude of coefficients for each feature. Features such as NumCompaniesWorked, YearsSinceLastPromotion, and OverTime_Yes have relatively higher positive coefficients, indicating their significant role in predicting attrition. Conversely, features like TotalWorkingYears, YearsInCurrentRole, OverTime_No, and JobSatisfaction have negative coefficients, suggesting that lower levels of satisfaction and involvement are associated with higher attrition rates.



6. **Handling Imbalanced Data with SMOTE**: To address the issue of class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) is employed to balance the dataset. A new SVM model is trained using the balanced data, and its

performance is evaluated. While SMOTE improves the recall for the minority class (attrition), it leads to a decrease in precision, indicating a trade-off between sensitivity and false positive rate.

Key Findings and Recommendations

- **Influential Factors**: The analysis revealed that factors such as job satisfaction, work-life balance, relationship satisfaction, and years at the company significantly impact employee attrition rates.
- **Predictive Model Performance**: The best-performing model achieved an accuracy of 89.9% and an F1-score of 0.519 on the test dataset, demonstrating its effectiveness in predicting employee attrition.
- **Retention Strategies**: Based on the findings, the following strategies are recommended to improve employee retention:
    - Enhance work-life balance initiatives, such as flexible schedules or remote work options.
    - Develop career development and mentorship programs to foster job satisfaction and growth opportunities.
    - Implement employee engagement surveys and feedback mechanisms to identify and address dissatisfaction promptly.
    - Offer competitive compensation and benefits packages to ensure employees feel valued and recognized.

Further Research and Limitations

While this project provides valuable insights and a predictive model for employee attrition, there are opportunities for further research and improvement, such as incorporating additional data sources, exploring advanced machine learning techniques, conducting longitudinal studies, and addressing potential biases or limitations in the dataset.

Conclusion

The project provides valuable insights into employee attrition and performance, offering actionable recommendations for organizations to enhance employee retention strategies and improve overall productivity.

Contributing

Contributions to this project are welcome. If you find any issues or have suggestions for improvements, please open an issue or submit a pull request.