Employee Attrition Data Analysis

GitHub Repo: https://github.com/SaivarunNamburi/Telecommunication-People-Data-Analysis

Introduction:

This documentation provides a comprehensive analysis of employee attrition at a large telecommunication company. The company faces an annual attrition rate of 15%, which leads to project delays, increased recruitment costs, and reduced productivity. The goal is to model the probability of attrition, identify key factors influencing it, and provide actionable recommendations to reduce the attrition rates.

Steps I've followed to achieve goals:

1. **Data Loading and Preparation:**
   - Loaded datasets: employee_survey_data.csv, general_data.csv, manager_survey_data.csv, in_time.csv, and out_time.csv.
   - Merged datasets based on EmployeeID to create a consolidated data frame.
   - Cleaned the data by handling missing values and encoding categorical variables.

   Missing Values:

	data
WorkLifeBalance	38
EnvironmentSatisfaction	25
JobSatisfaction	20
NumCompaniesWorked	19
TotalWorkingYears	9
Age	0
PercentSalaryHike	0

2. **Exploratory Data Analysis(EDA):**
   - Analyzed distribution and correlation of features
   - Visualized key statistics and distributions using bar plots and heatmaps
     - Correlation Analysis: Displays the correlation matrix using a heatmap to visualize the strengths between features
     - Cluster Analysis: Scatter plot of clusters formed by similar employee attributes, colored by attrition
     - Attrition Distribution by Key Feature: Bar plots showing the distribution of attrition across key features such 'JobSatisfaction', 'EnvironmentSatisfaction', 'YearsAtCompany'
3. **Feature Engineering:**
   - Created a new feature such as total working hours by calculating the difference between 'in_time' and 'out_time'
   - Transformed categorical variables into numerical format by using one-hot encoding

   We'll convert categorical variables into numeric representations using one-hot encoding for non-ordinal categorical variables and label encoding for ordinal categorical variables

Non-ordianal Variables:

1. Attrition : Our Target Varibale (Yes/No)
2. BusinessTravel
3. Department
4. EducationField
5. Gender
6. JobRole
7. MaritalStatus

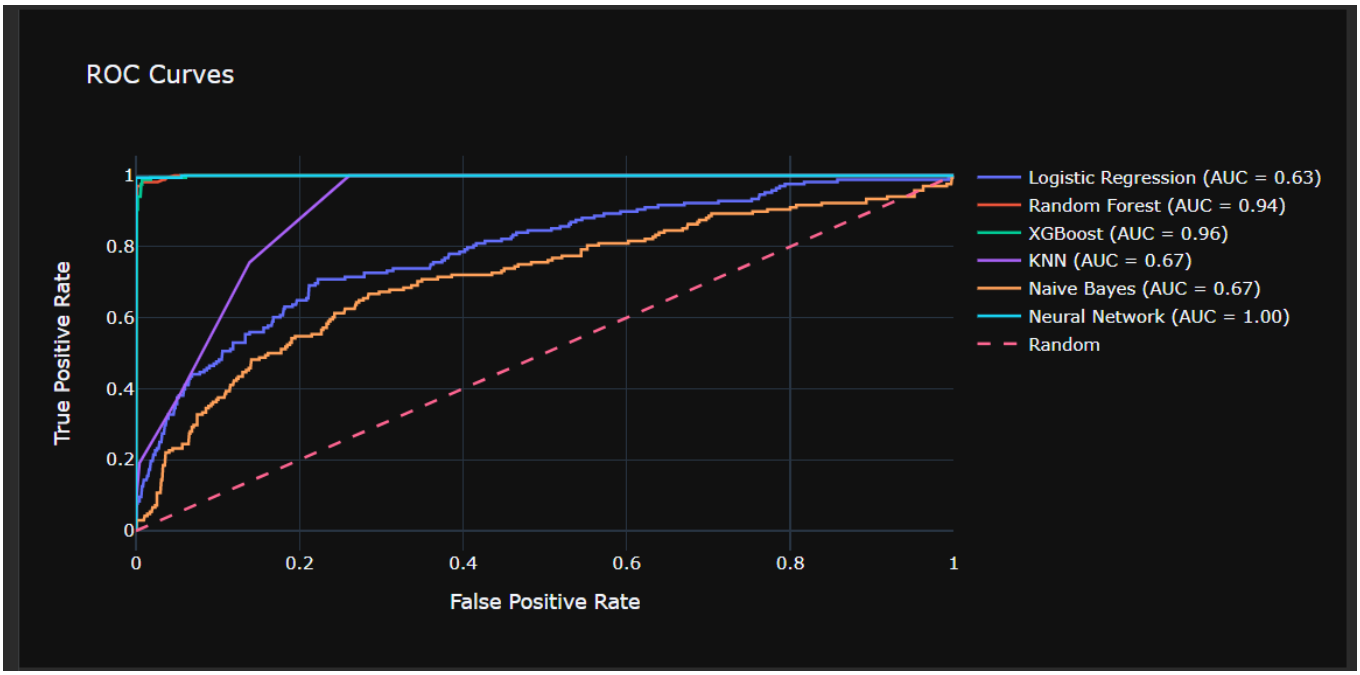
Ordinal Variables:

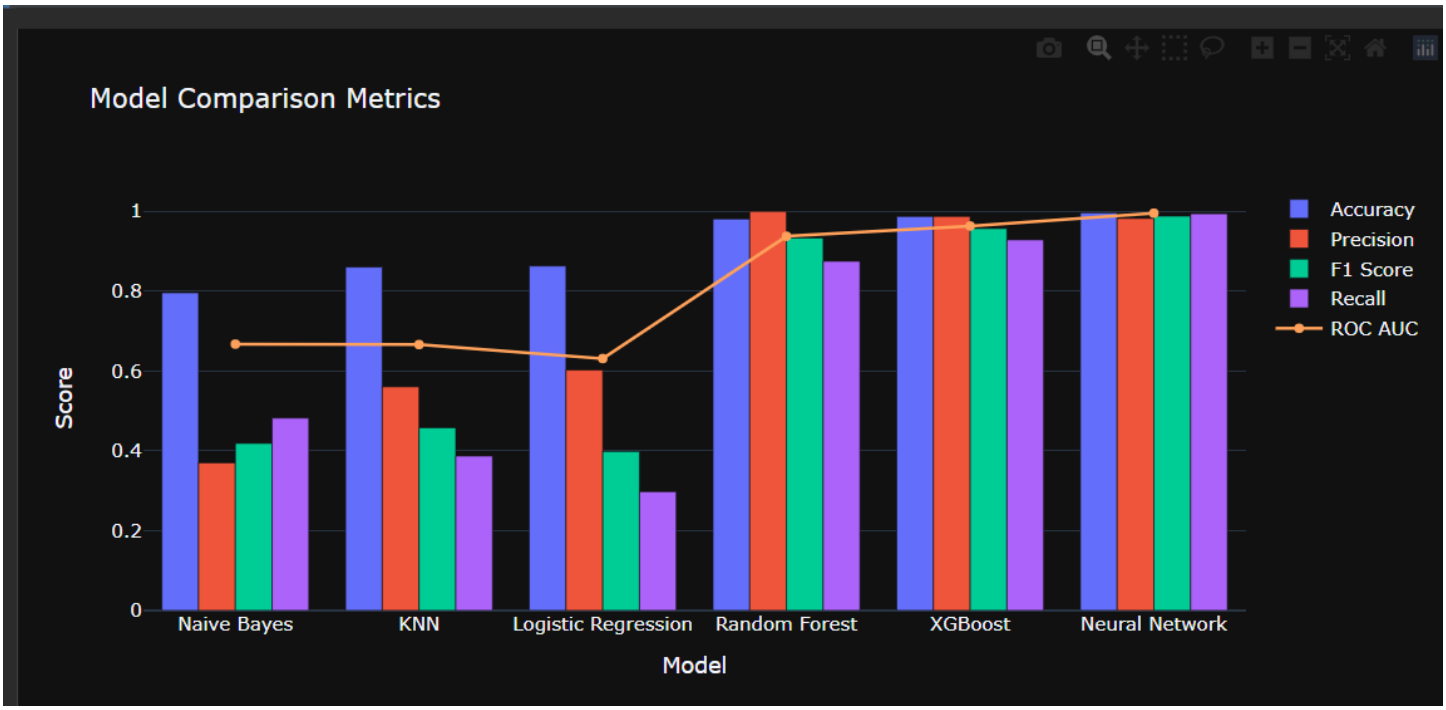
1. Education
2. EnvironmentSatisfaction
3. JobInvolvement
4. JobLevel
5. JobSatisfaction
6. PerformanceRating
7. StockOptimalLevel
8. WorkLifeBalance

Standardizing the features ensures that each feature has a mean of 0 and a standard deviation of 1, which can help with the convergence of the ML models.

4. **Modeling the Probability of Attrition:**
    - Split the data into training and testing sets.
    - Trained multiple models, including Logistic Regression, Neural Network, Random Forest, KNN, and XGBoost, to predict employee attrition.
    - Evaluated model performance using metrics like accuracy, precision, recall, F1 score, ROC AUC, and confusion matrix
    - Comparison of model performance using metrics like accuracy, precision, recall, F1 score, ROC AUC, and confusion matrix using plots
    - Extracted probabilities of attrition from the models.

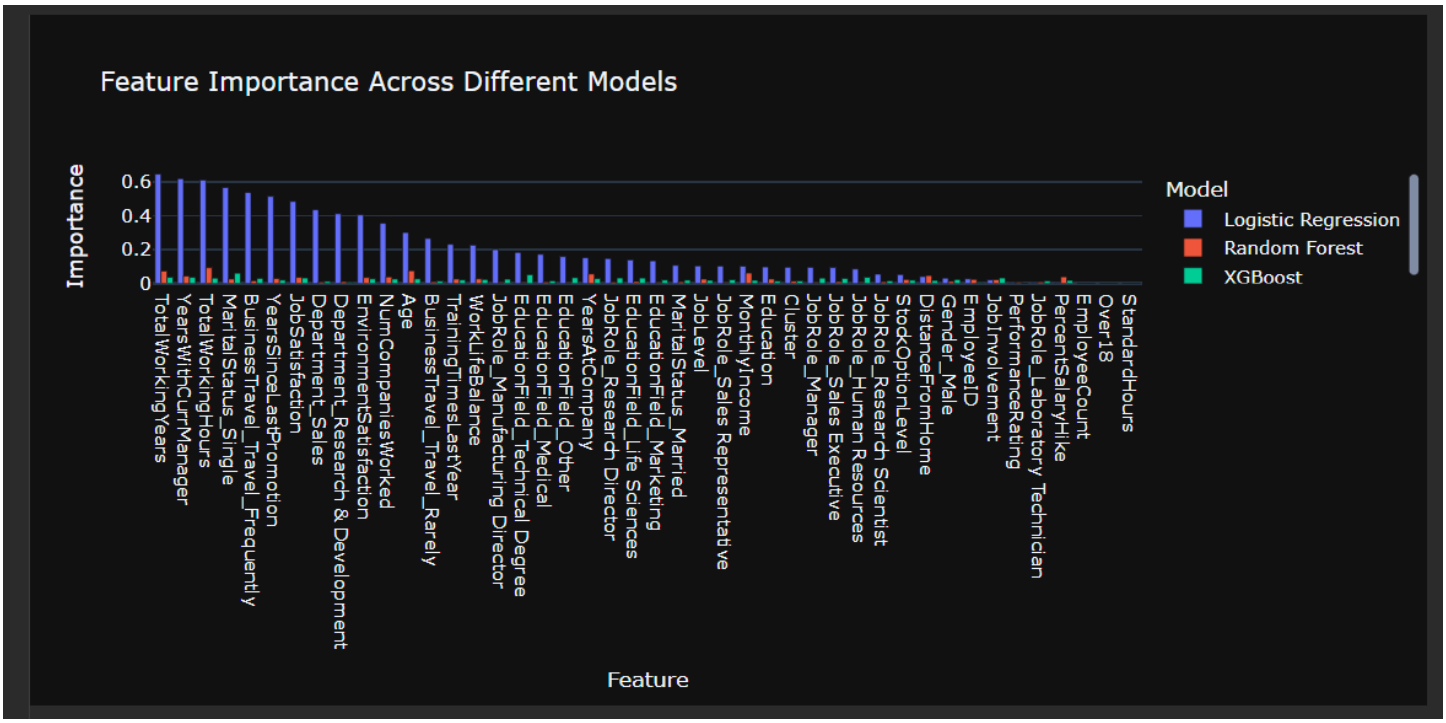
	Model	Accuracy	Precision	F1 Score	Recall	ROC AUC
5	Neural Network	0.996374	0.982353	0.988166	0.994048	0.995420
2	XGBoost	0.987307	0.987342	0.957055	0.928571	0.963216
1	Random Forest	0.980961	1.000000	0.933333	0.875000	0.937500
0	Logistic Regression	0.863101	0.602410	0.398406	0.297619	0.631162
3	KNN	0.860381	0.560345	0.457746	0.386905	0.666180
4	Naive Bayes	0.796011	0.369863	0.418605	0.482143	0.667275





5. **Feature Importance:**
    - Analyzed feature importance from different models to identify key factors influencing attrition.
    - Visualized feature importance using bar plots to identify which one is affecting most.

    The feature importance analysis provides insights into the key factors influencing attrition. By looking at the above graph we can say that "TotalWorkingYears", "YearsWithCurrManager", "BusinessTravel_Travel_Frequency", "MaritalStatus_Single", "YearsSinceLastPromotion" etc.. are significant predictors of attrition. Management should focus on these areas to improve retention



Focus on Employer Engagement

Total Working Years: Long-term employees may leave if they don't see further growth opportunities. Here management can implement career development programs and regular promotions.

Years With Current Employers: Employers with longer tenure with their current manager are less likely to leave. This implies the importance of managerial relationships in employee retention.

Business Travel Frequency: Employees who travel frequently for business are more likely to leave. Here management can provide better travel support or reduce travel could help in retaining these employees

Years Since Last Promotion: Employees who have not been promoted recently are more likely to leave. This highlights the importance of timely promotions.

Marital Status: Single employees are more prone to attrition, possibly due to fewer personalities. Developing engagement strategies targeted at single employees might help in reducing attrition.

6. **Key Insights and Actionable Recommendations:**
    - Key features influencing attrition were identified: Total Working Years, Years With Current Manager, BusinessTravel_Travel_Frequently, MaritalStatus_Single, YearsSinceLastPromotion.
    - Recommendations include enhancing managerial relationships, optimizing business travel policies, focusing on career progression, and improving work-life balance.

Enhance Managerial Relationships:

- Invest in leadership training to improve the quality of managerial relationships.
- Implement a mentorship program where experienced employees can guide newer ones.

Career Progression and Development:

- Establish clear career paths and communicate them to employees. Provide opportunities for continuous learning and development.

Optimize Business Travel Policies:

- Reduce the frequency of mandatory business travel, especially where remote communication can suffice.
- Offer flexible travel options and additional support(like reimbursement, and company vehicle) for employers who must travel frequently.

Recognition and Rewards:

- Provide meaningful rewards and incentives based on performance and tenure. Celebrate employee achievements and contributions.

Work-Life Balance:

- Promote policies that support a healthy work-life balance, such as flexible working hours and remote work options.
- Conduct regular surveys of employer satisfaction with work-life balance and make necessary adjustments.
- Organize social events, team-building activities, and clubs to improve social connections within the workplace.

7. Recommendations for Data Collection/Analysis Process

Data Collection

- **Automate Data Collection**: Implement automated data collection methods to reduce errors and ensure real-time data availability.
- **Integrate Systems**: Ensure all data systems are integrated, allowing for seamless data flow and reducing the need for manual data consolidation.
- **Standardize Data Entry**: Establish standardized data entry protocols to maintain consistency and accuracy across datasets.

Data Analysis

- **Regularly Update Models**: Frequently update predictive models with new data to ensure they remain accurate and relevant.
- **Conduct Regular Training**: Ensure that data analysts and other relevant staff receive regular training on the latest data analysis methodologies and tools.
- **Implement Data Quality Checks**: Regularly perform data quality checks to identify and rectify any inconsistencies or errors in the data.

8. Conclusion:

- By implementing the above recommendations, we can address the factors influencing employee attrition and improve retention rates. The use of advanced models like Neural Network, or XGBoost provides accurate predictions and actionable insights, enabling data-driven decision-making to create a better work environment and reduce attrition.