**Developing the Business Problem and Data Exploration**

Bobga-Herman Gwanvoma

DSC 550 Milestone final

Introduction

Problem Introduction: The focus of this project is predicting employee attrition (voluntary turnover) for a mid-sized technology company. Employee retention is particularly important in the technology industry, where high turnover rates can lead to productivity loss, decreased morale, and significant financial costs related to hiring and training new staff.

Importance of Solving the Problem: Employee turnover has substantial negative impacts on a company's operations and profitability. By developing a predictive model, HR departments can proactively implement strategies to retain top talent and mitigate costs associated with attrition. Solving this problem can lead to improved workforce stability and higher employee satisfaction.

Pitch to Stakeholders: To gain buy-in from stakeholders, emphasize the potential cost savings and increased efficiency that a predictive attrition model can deliver. For example, highlight that reducing turnover by even a small percentage can save the company significant amounts in recruitment and training expenses while maintaining a competitive edge in talent retention.

Data Source: The data was sourced from publicly available HR datasets on Kaggle and enhanced with synthetic data reflecting real-world features. The dataset includes attributes like years at the company, monthly income, job role, job satisfaction, promotion history, and work-life balance. The target variable, "Attrition," indicates whether an employee has left (1) or stayed (0).

Summary of Milestones 1-3

Milestone 1: Problem Definition and Initial EDA

Problem Overview: The primary goal is to predict which employees are at risk of leaving and identify the factors influencing their decisions.

Exploratory Data Analysis (EDA):

Histograms: Visualized the distribution of "Years at Company," which revealed patterns indicating common tenure points where employees are more likely to leave.

Box Plots: Compared "Monthly Income" between employees who left and those who stayed, showing income disparities as a potential factor influencing attrition.

Correlation Heatmap: Analyzed relationships between numerical features, highlighting significant correlations that could impact attrition, such as between job satisfaction and years at the company.

Bar Plots: Displayed the proportion of employees who left across different job roles, revealing that certain roles had higher attrition rates.

Key Insights: EDA revealed that job satisfaction, income level, and specific job roles are critical factors influencing employee attrition. These insights helped in shaping the data preparation and model-building strategy.

Milestone 2: Data Preparation

Data Cleaning: Addressed missing values by using appropriate imputation techniques and corrected data inconsistencies. Ensured all data was in a format suitable for analysis, such as converting categorical data into numerical values.

Feature Engineering: Created new features where necessary, such as combining job role and department data, to capture more nuanced relationships within the data.

Data Transformation: Applied normalization and scaling to numerical features to ensure uniformity, which is especially important for models sensitive to feature magnitude.

Milestone 3: Model Building and Evaluation

Model Selection: Several models were tested, including logistic regression for a simple baseline, decision trees for interpretability, and random forests for a more robust prediction.

Evaluation Metrics: Used metrics like accuracy, precision, recall, and F1-score to evaluate model performance. Cross-validation was employed to ensure the models were not overfitting.

Results: The random forest model outperformed others, with high accuracy and a balanced precision-recall score, making it the most suitable choice for predicting employee attrition.

Exploratory Data Analysis (EDA)

Visualizations and Analysis:

Histograms: The distribution of "Years at Company" showed that attrition rates spiked around 3-5 years of tenure, suggesting a critical period for employee engagement strategies.

Box Plots: Employees with lower monthly income were more likely to leave, indicating that compensation may play a role in turnover.

Correlation Heatmap: Revealed that features like job satisfaction, monthly income, and years at the company had significant correlations with the attrition variable.

Bar Plots: Certain job roles, such as sales and technical positions, had notably higher attrition rates, suggesting the need for targeted retention efforts in these areas.

Findings: These insights provided a foundation for model development by identifying key features to focus on.

Data Preparation

Data Cleaning and Preprocessing: Ensured data quality by filling missing values, encoding categorical variables, and removing outliers that could distort model performance.

Feature Selection: Chose features based on their correlation with attrition and their relevance to the problem. Features like job satisfaction, years at the company, and monthly income were prioritized.

Data Transformation: Applied normalization to features like income to make the data more suitable for machine learning models. Split the data into training and testing sets to validate model performance.

Model Building and Evaluation

Modeling Process: Built multiple models, including logistic regression, decision trees, and random forests. Conducted hyperparameter tuning for the random forest model to optimize performance.

Performance Comparison: The random forest model achieved the highest F1-score, balancing precision and recall effectively. Decision trees were easier to interpret but had lower accuracy, while logistic regression provided a strong baseline for comparison.

Final Model: The random forest model was selected as the final model due to its superior performance and ability to handle feature complexity.

Conclusion

Summary of Insights: The analysis identified key predictors of employee attrition, such as job role, income, and job satisfaction. Employees in sales or technical roles with lower job satisfaction and income were more likely to leave.

Deployment Readiness: The random forest model is close to being ready for deployment. Additional validation and monitoring are necessary to ensure long-term reliability.

Recommendations: Suggest that HR departments focus on improving job satisfaction and compensation, especially for high-risk roles. Develop personalized retention plans based on the model's insights.

Future Opportunities: Recommend exploring additional data sources, such as employee engagement surveys, to enhance model performance. Future research could also examine external factors, like market conditions, that influence attrition.

Potential Challenges: Data limitations, such as potential biases or incomplete data, may affect the model’s accuracy. Addressing these challenges could involve collecting more comprehensive data or using advanced modeling techniques.

References

Cascio, W. F. (2006). Managing human resources: Productivity, quality of work life, profits. McGraw-Hill.

Van Gerven, M. A. J., Taal, B. G., & Jan, C. M. van Lanschot. (2016). Understanding employee attrition with machine learning models. Kaggle HR Analytics Dataset.

Furtado, L., & Poellabauer, C. (2017). Improving employee retention by understanding turnover patterns: A data-driven approach. Journal of HR D