**DSC 550 Term Project Final Submission**

Christina Vosnak

DSC 550

Professor Werner

November 17th, 2024

**Introduction**

In the field of human resources, particularly within Talent Acquisition, one of the most pressing questions is understanding why employees choose to leave a company. High turnover rates often signal potential issues related to workplace culture, management practices, and overall employee satisfaction. These factors can harm a company’s reputation, increase recruitment costs, and deter potential candidates from applying for open positions. Understanding the reasons behind employee turnover is essential for human resources teams to implement effective retention strategies that foster a more stable and productive workforce, helping companies maintain a competitive edge in attracting and retaining top talent. By researching the underlying factors driving turnover, companies can make positive changes to their culture, empowering managers to connect more effectively with their direct reports and creating a more engaging work environment. Addressing turnover not only improves the company’s external reputation by offering potential employees a positive view of its culture but also allows organizations to retain skilled employees, reduce hiring costs, and build long-term value for both employees and the organization.

To gain stakeholder buy-in, I would emphasize both the strategic and financial benefits of addressing employee turnover, focusing on the measurable differences in employee satisfaction, recruitment costs, and productivity between companies with low turnover and those with high turnover. By presenting data on the company’s current turnover rates, recruitment expenses, and productivity levels, I would draw a clear comparison that highlights the negative impacts of high turnover and the potential opportunities for improvement.

I would also suggest actionable strategies to decrease turnover and recruitment costs while increasing team morale and employee satisfaction. Framing this initiative as a way to enhance both the company’s financial performance and its workplace culture, I would demonstrate how investing in turnover reduction strategies could lead to long-term benefits, including a more stable, engaged workforce and a stronger reputation as an employer of choice.

For this project, I obtained my data from Kaggle, a widely recognized platform for data science and machine learning projects. It is important to note that the dataset used for this analysis contains entirely fictional information, including detailed data on employee attrition, job satisfaction, and performance. Despite being synthetic, Kaggle’s datasets are curated to reflect realistic patterns and trends across diverse industries, making them a reliable resource for developing actionable strategies. By leveraging this dataset, I conducted an in-depth analysis that provides valuable insights, which can be directly applied to inform and enhance retention strategies within real-world organizations.

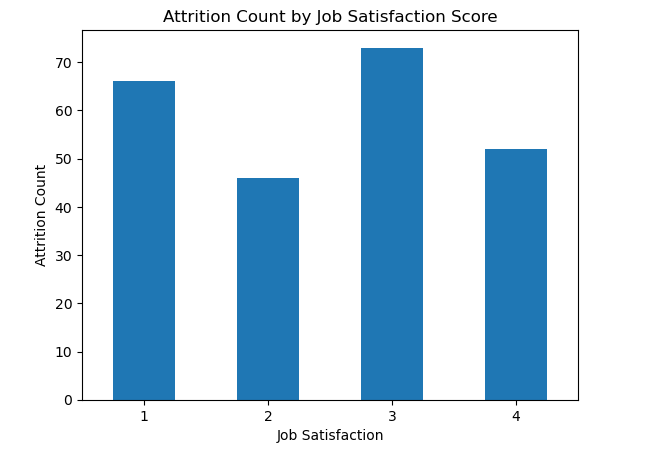
**Milestone Summaries**

Milestone 1 of this project centered on exploring the dataset in its rawest form. Without making any manipulations, I focused on analyzing the initial structure and characteristics of the data, creating a series of visualizations to uncover key patterns and trends. These visuals provided a foundational understanding of the dataset and offered preliminary insights to begin addressing the primary question behind this project: factors that contribute to employee attrition.

The first visual examines the Attrition Count by Department, providing an initial overview of how employee turnover is distributed across the organization. This visual immediately highlights an imbalance in the dataset between employees who left the company and those who stayed. Notably, the Research & Development department shows the highest counts of attrition. However, it is important to consider that this department also has the largest number of employees among the three teams, which may explain the higher attrition numbers. This finding underscores the need to investigate attrition rates relative to department size to better understand trends and identify areas of concern.

A graph of a bar chart

Description automatically generated

The second visual examines **Attrition Counts by Job Satisfaction Score**, offering insight into how job satisfaction levels might relate to employee turnover. Surprisingly, the chart reveals no clear correlation between these two variables, as the distribution of attrition rates appears relatively consistent across all satisfaction scores. This lack of significant variation suggests that job satisfaction, at least as captured in this dataset, may not be a strong predictor of employee attrition.

A graph of a number of bars

Description automatically generated The third visual examines Attrition Count by Age, highlighting the distribution of employee ages across the company and how they relate to turnover. While no strong correlation between attrition and age is immediately evident, the chart reveals notable trends. Attrition counts begin to rise at age 24, peak at ages 29 and 31, and then decrease, stabilizing at consistent levels starting at age 36. These trends may suggest that employees in their late 20s and early 30s are more likely to leave the company, possibly due to career transitions or other life changes common at this stage.

The fourth visual looks at Total Working Years compared to Monthly Income and puts the data into scatterplot form to visualize the relationship between Total Working Years and Monthly Income, with data points differentiated by attrition status (red = "Yes," black = "No"). The chart shows that as total working years increase, monthly income generally trends upward, reflecting expected career progression. However, it is notable that attrition (indicated by red dots) occurs across various levels of income and experience, with no clear boundary. There seems to be a concentration of attrition among employees with fewer than 20 years of total working experience and those that earn less than $10,000 per month, suggesting that this group might be more vulnerable to turnover. Additionally, some attrition is observed in higher income brackets, indicating that income alone may not fully prevent employees from leaving the company.

A graph of red and black dots

Description automatically generated

A graph of income groups

Description automatically generated with medium confidence The final visualization illustrates Attrition Counts by Income Group, categorizing employees into income ranges and visualizing how many left the company. The chart reveals a clear trend: employees earning less than $4,000 per month have the highest attrition count by a significant margin. Attrition decreases as income increases, with notably fewer employees leaving in the higher income brackets ($12,000+). This pattern suggests that income level may play a significant role in employee retention, with lower-income employees being more likely to leave the organization.

Overall, this initial exploration serves as a critical starting point, setting the stage for more detailed analysis and data preparation in the next phases of the project. By understanding the dataset's composition and identifying potential patterns, I can begin to formulate strategies and methodologies to address the underlying causes of attrition effectively. This milestone helped to lay the groundwork for deeper insights and actionable recommendations in subsequent steps regarding data preparation and model building.

Milestone 2 of this project focused on data preparation, a crucial step in transforming the raw dataset into a format suitable for detailed analysis and modeling. This phase involved cleaning the data, addressing any inconsistencies or missing values, and ensuring the dataset was ready for further exploration. Additionally, relevant variables were identified, and data transformations were performed to enhance interpretability and analytical accuracy.

Several features were dropped during this stage due to their lack of relevance or interpretability in addressing employee attrition. Employee Count, Over18 and Standard Hours were three columns that contained identical values across all rows. Since these columns provide no variability, they offer no insight for the analysis or modeling. Additionally, Hourly Rate, Daily Rate, and Monthly Rate, were dropped as they were columns that attempted to standardize hourly and salaried employee earnings but without any accompanying framework regarding hours worked. Instead, I retained the Monthly Income column to serve as a more meaningful and interpretable metric of employee compensation. Luckily, this dataset had no missing values therefore no modification of the data was necessary at this stage.

Additionally, two new features were engineered to enhance the analysis. This includes the Income Group column which was created during Milestone 1 to categorize employee’s monthly income into $4,000 buckets to make it easier to identify trends across income levels. This also includes the Overall Satisfaction column which combines satisfaction metrics from the Work Life Balance, Relationship Satisfaction, Job Satisfaction, and Environment Satisfaction columns into a single score. By aggregating these metrics, the feature allows for a holistic view of employee satisfaction and its potential effect on attrition. I also created dummy variables by transforming categorical variables to prepare the data for machine learning models. This transformation ensured the dataset was in a numerical format suitable for analysis while avoiding multicollinearity by dropping one category from each categorical variable.

Finally, I explored two methods of dimensionality reduction to simplify the dataset. The first is a Principal Component Analysis that maintains 95% variance which reduced the number of features from 47 to 33. This technique ensures most of the data’s variance is retained while reducing complexity. The second is a High Variance Thresholder that used a variance threshold of 95% to reduce the features from 47 to 18 retaining only those with the highest variability. This step provides a preliminary view of feature importance and helps identify potential redundancies.

Overall, this milestone established a clean, transformed dataset ready for further analysis and modeling. Engineered features and reduced dimensionality provide a strong foundation for identifying patterns and building effective predictive models in the next milestone which will involve incorporating strategies to handle data imbalance, and test models to evaluate which yields the most accurate results.

Milestone 3 of this project focused on building, testing, and evaluating the accuracy of multiple predictive models. For this analysis, I compared three models: Logistic Regression, Decision Tree Classifier, and Random Forest Classifier. Each model was trained and tested across three versions of the dataset: the original dataset with all features included, the version with Principal Component Analysis (PCA) applied, and the version with High Variance Threshold applied. Model performance was assessed using a range of metrics, including accuracy, precision, recall, F1 score, AUC-ROC, and confusion matrices. To gain deeper insights into the models' results, I identified the top five most important features for each model using the original dataset, providing actionable recommendations on which variables played the most significant role in predicting employee attrition. The results are illustrated below.

**A screenshot of a computer

Description automatically generated**

**A black and white text with black letters

Description automatically generated with medium confidence**

**A screenshot of a computer

Description automatically generated**

**A black text on a white background

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A number of numbers and symbols

Description automatically generated with medium confidence**

The logistic regression with all features included reported accuracy of 79% with strong precision of 84%, recall at 79.5% and f1 score of 81%. The AUC-ROC was 83.3%, indicating good discriminatory ability. The confusion matrix shows a balanced classification of attrition and non-attrition cases. Compared to PCA, the accuracy dropped slightly to 77.6%, precision dropped to 83.67%, recall dropped to 77.55% and f1 score also dropped to 79.4%. However, the AUC-ROC increased slightly to 84.4%. Moving on to the High Variance Threshold, this version performed significantly worse, with accuracy falling to 63.6%, precision falling to 75.4%, recall falling to 67.5%, f1 score falling to 67.5% and the AUC-ROC falling to 65.5%. Overall, the confusion matrices also saw a decrease in accuracy from version with all features to the PCA data to the High Variance data. The top 5 features were OverTime (Yes), PerformanceRating, IncomeGroup, BusinessTravel, and MaritalStatus.

The decision tree classifier with all features included reported accuracy of 80.9% with strong precision of 79.5%, recall of 80.9%, and an f1 score of 81.1%. The AUC-ROC was low at 64%, indicating weaker discriminatory ability. Compared to PCA, the accuracy increased to 81.2&, precision decreased to 78.7%, recall increased to 81.2% and f1 score decreased to 79.6%. The AUC-ROC also decreased to 61.3%. Moving on to the High Variance Threshold, this version performed significantly worse again reporting accuracy at 73.8%, precision at 74.3%, recall at 73.8%, and an f1 score at 74%. The AUC\_ROC significantly dropped to 56.7%, once again indicating an even weaker discriminatory ability. However, the confusion matrices for this model all reported significantly better results than those from the logistic regression model. The top features from this model include OverTime (Yes), YearsAtCompany, Age, EmployeeNumber, and DistanceFromHome.

The random forest classifier with all features included reported accuracy of 83.6%, precision of 81.6%, recall of 83.67%, f1 score of 78.3% and AUC-ROC at 77.4%. Compared to PCA, the accuracy is 84%, precision is 86%, recall is 84%, f1 score is 78%, and AUC-ROC is 81%. Moving on to the High Variance Threshold, the accuracy is 81.9%, precision score is 76.9%, recall is 81.9%, f1 score is 76% and the AUC is 66%. The confusion matrices for all of these versions are significantly better than both the decision tree classifier and logistic regression, reporting the highest true negative and true positive results. The top five features from this model include MonthlyIncome, OverTime (Yes), Age, YearsAtCompany, and EmployeeNumber.

**Conclusion**

Overall, the best-performing model based on the above analysis is the Random Forest Classifier using all features, as it consistently achieved the highest metrics across all evaluation criteria. Among the top five features for this model, Monthly Income was the most significant, with an importance score of 0.0776. However, a notable observation is that OverTime\_Yes appeared in the top five features for all three models, highlighting its consistent relevance in predicting employee attrition.

Across all models, datasets with all features included performed the best, providing the most accurate and reliable predictions. Models using PCA-reduced datasets performed slightly worse but still delivered reasonable results. In contrast, models with the High Variance Threshold dataset performed significantly worse, likely due to the excessive reduction of features, which removed critical variables necessary for accurate predictions.

At this stage, while the Random Forest Classifier model with all features produced strong results, I do not believe it is ready for deployment. The model would benefit from adjustments to better address the class imbalance in the dataset, which could potentially improve its performance from the high 80s to the low 90s in terms of accuracy. Additionally, it is essential to test the model on new, unseen data to verify its ability to generalize and maintain consistent performance in real-world scenarios. I recommend that the model itself includes a SMOTE or threshold tuning technique to handle imbalances within the data. I also recommend this model be tested significantly more to yield more results.

Regarding the results of this model I’d advise the stakeholders to partner with Human Resources Business Partners to reevaluate position wages to ensure they are relative to averages across the industry. Stakeholders should also reevaluate their overtime policies as the model shows that employees who work overtime are more likely to leave the company.

Based on the results of the model, I recommend that stakeholders collaborate closely with Human Resources Business Partners to address two key areas: compensation and overtime policies. First, the organization should reevaluate position wages to ensure they are competitive with industry averages. Compensation emerged as one of the most significant factors influencing employee attrition, with lower-income employees being more likely to leave. Conducting a comprehensive market analysis to benchmark salaries against competitors and implementing adjustments where necessary could help improve employee satisfaction and retention.

Second, stakeholders should thoroughly review and potentially revise their overtime policies. The model consistently highlighted OverTime\_Yes as a top predictor of attrition, suggesting that employees who work overtime are at a higher risk of leaving the company. This may be due to factors like burnout, poor work-life balance, or feelings of being overworked. To address this, the organization should explore strategies such as capping overtime hours, offering more flexible scheduling, or providing additional support and recognition for employees who regularly work overtime. Ensuring that employees feel valued and supported in their roles could significantly mitigate the negative impact of overtime on attrition. Together, these steps could lead to a more satisfied and stable workforce, ultimately reducing turnover and supporting the organization’s long-term goals.

While the current analysis and model development provide valuable insights, several challenges and opportunities remain to be explored. One potential challenge is that further exploration is needed to assess the impact of external factors not included in the dataset, such as market trends, industry benchmarks, or economic conditions, which could influence attrition. Opportunities also exist to refine the feature set by engineering more nuanced variables, such as tracking changes in job satisfaction or overtime patterns over time, to better capture dynamic predictors of attrition. Finally, expanding the scope of the analysis to incorporate qualitative data, such as employee feedback from surveys or exit interviews, could provide a richer context for understanding the underlying reasons behind turnover. These efforts would not only enhance the model’s predictive capabilities but also offer deeper, actionable insights for reducing attrition and improving employee retention.