Predicting the Risk of Employee Attrition

ISYE 6740

Ravi Sunder, Noumik Thadani
Group 75

Contents

1.1 Problem Statement

Attrition Epidemic

Across the globe, companies are facing an attrition epidemic. The attrition rate, or labor turnover rate, is the rate at which employees leave a company either voluntarily or involuntarily. Companies initiate retention policies to compel employees to continue working at the company for longer durations to save the company money on recruitment. This project aims to build an attrition prediction model to proactively identify employees at risk of leaving. Having a better understanding of factors leading to attrition would allow a company to reduce recruitment costs by improving retention rates, implementing effective retention strategies, and developing more proactive workforce planning methods.

Cost of Attrition

Attrition is expensive. When hiring employees, companies incur large recruitment and training expenses, as well as lost productivity of the prior employee. By predicting attrition, organizations can allocate resources more effectively and budget to minimize recruitment costs. According to the Society for Human Resource Management (SHRM) Human Capital benchmarking report, the average cost per hire has risen 14% from 2019 to 2023. Internal recruiting costs on average between $10,000 and $15,000 per employee and companies often budget between 15 and 25% of an employee's annual salary for external recruiting costs [1]. Consider a company of 1,000 employees with a relatively low yearly attrition rate of 10% [2]. This company would plan to spend between $1 and $1.5 million on in-house recruiting costs, before considering the cost of employing and maintaining a recruitment team. If the company utilized an external recruiting company, assuming the average employee earns $100k annually, the recruitment costs would increase to between $1.5 and $2.5 million. These expenses can take away from a company's bottom line and are only increasing year over year.

Retention Strategies and Challenges

Retention philosophy may be simple or complex depending on the industry and workplace environment, yet it is a complicated picture to understand without substantial evidence. For a company, the human resource department is tasked with developing retention strategies to save the company money. It can take considerable time to develop these strategies which begin with analysis of the data, and the data may be complex and difficult to analyze with simple methods. Additionally, organizations need to ensure they have the right talent in the right roles. Predictive models can help HR departments anticipate attrition trends and adjust budgeting and hiring plans accordingly.

Benefits of an Attrition Prediction Model

There are many potential benefits and use cases of having a model to predict attrition. Developing a robust retention strategy or flagging specific employees that may be more likely to leave the company are a few examples that we plan to explore. Firstly, the HR team can understand the variables that lead to an employee leaving the company, and thus target these variables as part of their retention strategy. Additionally, an attrition model could act as an early warning system, flagging employees at risk of leaving. HR teams can then intervene promptly by addressing concerns and improving job satisfaction. By analyzing historical data, the model can identify patterns associated with attrition. Companies could be able to understand how external (PESTEL) impacts may affect an employee's decision to stay with or leave the company. Furthermore, companies can tailor interventions based on individual employee profiles and address an employee's potential reasons for leaving. If an employee does leave, a prediction model can save time in identifying suitable successors to ensure a smoother transition. In conclusion, having an attrition prediction model will save a company time in developing retention strategies and money in reducing recruitment costs. Companies can gain a better understanding of their employees to enhance their job satisfaction, making the company a more attractive employer within the local community.

Drawbacks and Limitations

While we highlighted various benefits of having an attrition model, there are also numerous drawbacks to having such a model. An individual's decision to leave a job is complex and influenced by numerous human and external factors. Human factors include variables such as life events, relationships with coworkers, and job satisfaction, and external factors can include the season, economy, and political situation. It is important to recognize that data and models may struggle to capture these nuances. If employees are aware their companies are utilizing a model to predict attrition risk, they may disengage or seek other opportunities, exacerbating the problem. If the attrition model is overfit to the provided training data, it may not be useful for the company as a representation of their employee's attrition risk. Relying heavily on historical data may also lead to inaccurate predictions, especially if unforeseen changes occur.

2.1 Methodology

2.1.1 Overview

The approach to this project will be broken down as follows:

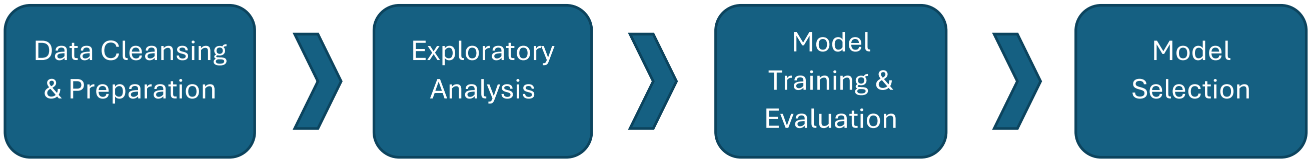

Figure 1: High-level Project Approach

2.1.2 Dataset

The data for this project is sourced from Kaggle. The dataset is "Employee Attrition Classification Dataset" [3]. It is important to note that the dataset is synthetic.

The dataset contains several quantitative features related to each employee, such as age, income, years at company, distance from home. It also contains qualitative features such as perceived work-life balance and perceived job satisfaction. The dataset contains features that are representative of the individual and features that represent the company the individual has worked for or are currently working for. The dataset contains 74,498 observations, with a provided split of 80% for training and 20% for testing.

2.1.3 Data Cleaning & Preparation

In this first step, the team analyzed each feature for missing data. The dataset did not have any missing data, potentially due to it being a synthesized dataset. All features were kept within the dataset for analysis, except for the 'Employee ID' column, which is a categorical identifier and not indicative of the individual's decision to remain at or leave their company. Additionally, due to the dataset being synthesized, the team was able to skip much of the data cleaning and preparation process to focus more on the model training and evaluation.

The team first converted all categorical columns into dummy variables, then the team scaled the data. After creating dummy columns, the dataset increased from 22 predictor variables to 41. The team was interested in differences in model performance between the full dataset and a dimensionality reduced version of the data. For dimensionality reduction, basic PCA was conducted. To choose the appropriate number of components in PCA, the team evaluated the explained variance vs number of components, seen in Figure 2. From this, the team opted for 28 principal components, which explained 86.3% of the total variance within the data. The team decided that 28 components was a good balance between dimensionality reduction and retaining features of the original data.

The team wanted to explore kernel based PCA. However, due to memory constraints, it was not possible to conduct. In Python's Sci-kit Learn library, kernel-based PCA requires the program create the full kernel matrix which is $O(n^2)$, and required 41.4 GB of RAM, which the team did not have available. However, through weighted sampling a kernel based PCA would be possible.

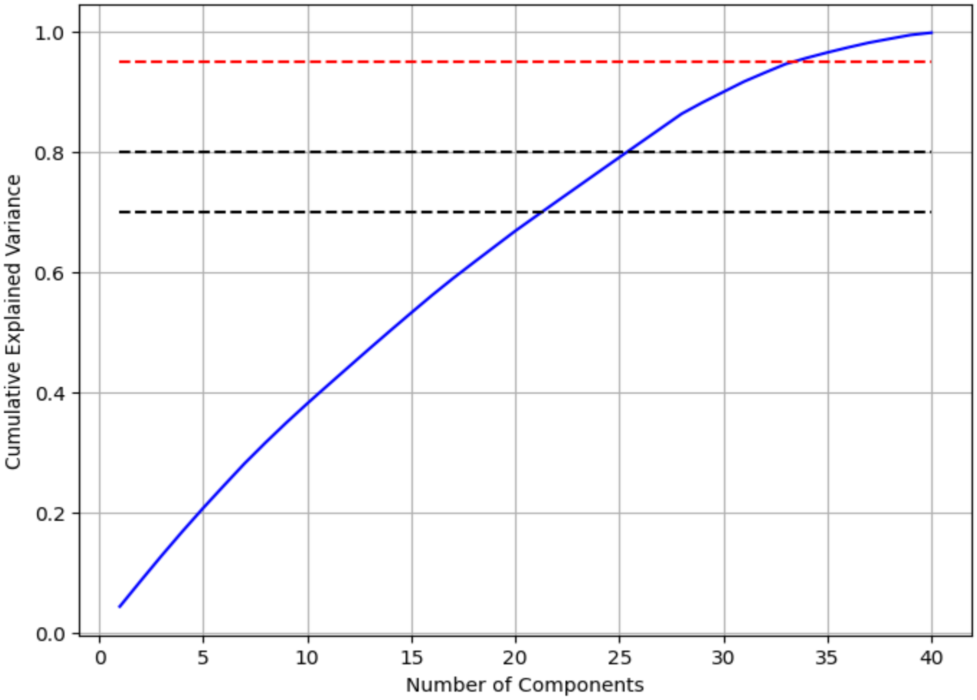

Figure 2: Total Explained Variance vs Number of Components for PCA

2.1.4 Exploratory Analysis

During exploratory analysis, the team gained an understanding of the prior distributions of the data. The team noticed that the provided split from Kaggle had equal distributions in their feature data. This can be difficult to achieve with random sampling. Ideally, the team would have liked to train models on the prepared Kaggle split and an additional random split to compare results. However, due to time constraints the team decided to move forward with just the Kaggle split. The advantage of using the Kaggle split is that the testing set is a good representation of the training set, so the validation results are a reliable representation of the model.

To visualize the high-dimensional data, the team performed t-SNE in both 2 and 3 dimensions for the PCA and non-PCA data. The 2D t-SNE, Figure 3.a and 3.b, shows that the data appears difficult to clearly separate between the two classes. The 3D t-SNE visualization is not included in this paper due to the lack of clarity in class separation in the three-dimensional space. The 3-D t-SNE were generally mixed in most areas, with very small clusters of each class observed throughout the mixture.

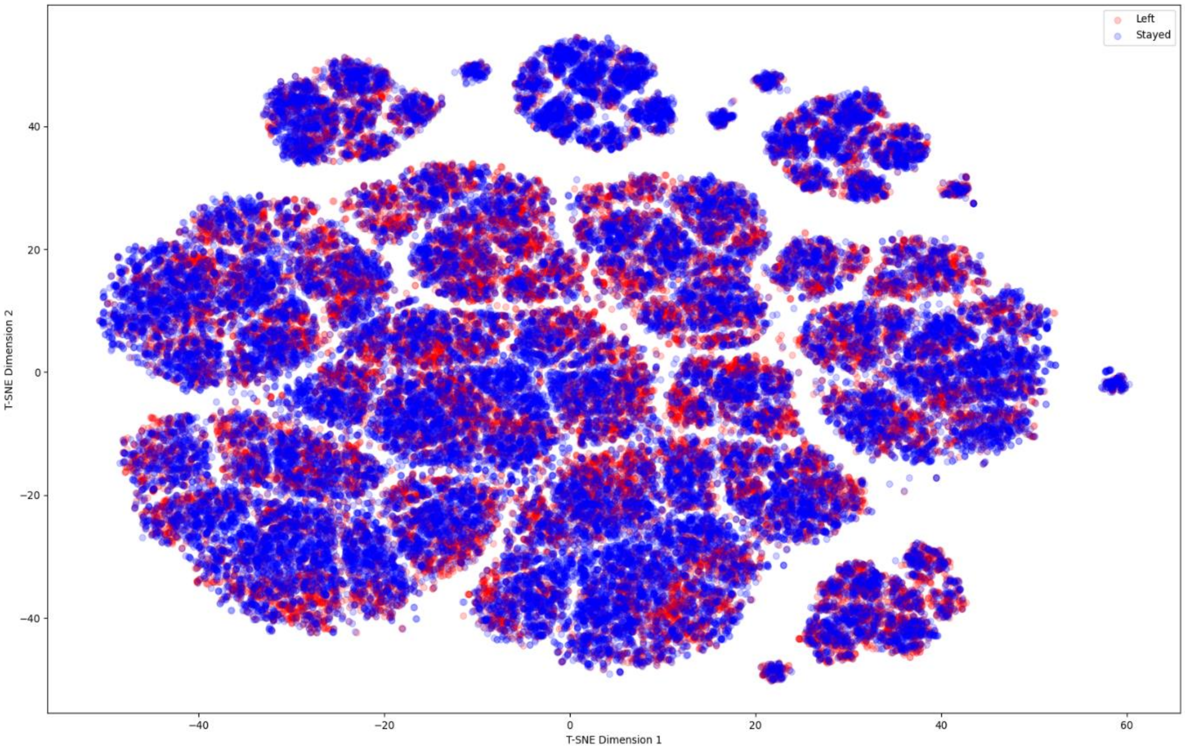

Figure 3.a, 2-D t-SNE Representation of Attrition Data, no PCA

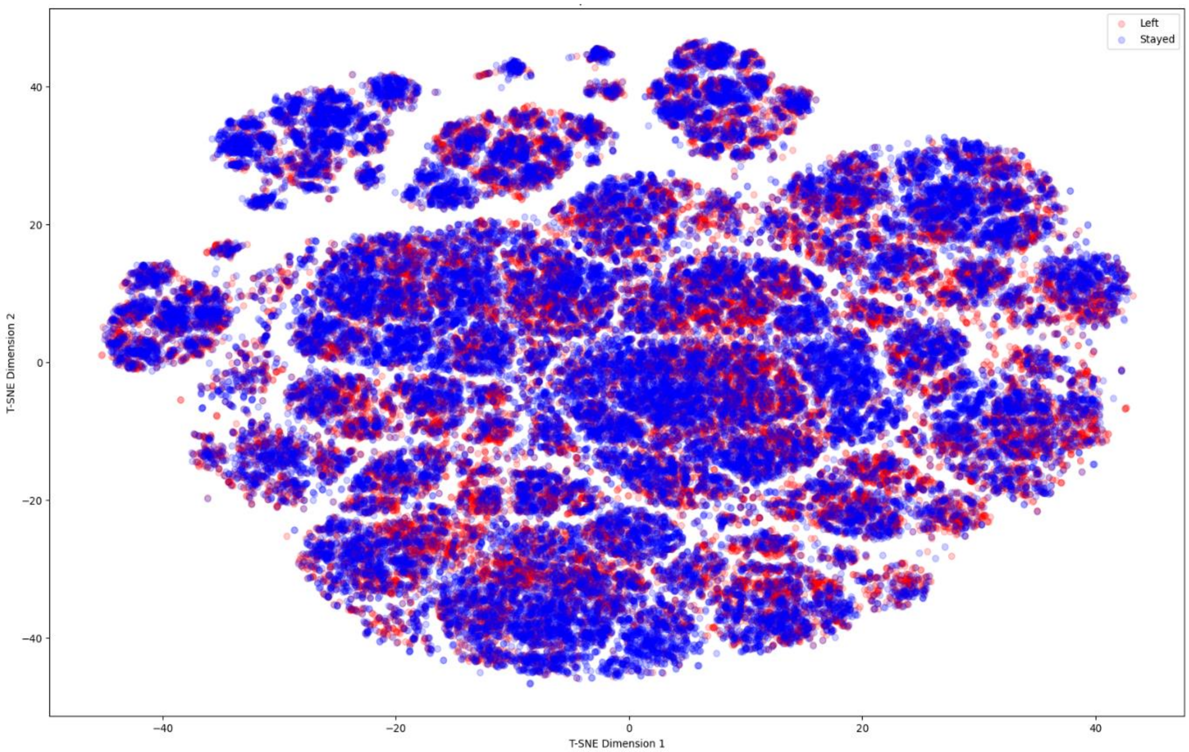

Figure 3.b, 2-D t-SNE Representation of Attrition Data, with PCA (n=28)

2.1.5 Model Training

The team attempted to train as many classification models as possible to compare the results. Models were trained on the testing data and evaluated on the training data provided by Kaggle. For most model training, grid search with 5-fold cross-validation was utilized to determine the optimal parameters for certain models. The team will touch more on this later, as the parameters that returned the highest cross-validation score were used, therefore the chosen models likely had a poor bias-variance tradeoff. This could potentially result in models over-fitting to the training data and producing poor results on the validation set, hampering model selection.

K-Nearest Neighbors (KNN)

For the KNN model, the team experimented with different values of k to find the optimal number of neighbors. The optimal k-values were chosen from cross-validation results, with k=491 and k=226 for data without PCA and with PCA, respectively. For the PCA data, the optimal k value is relatively close to the square root of the size of the training set, which is 244.1, however the optimal k value for non-PCA data is about double. Cross-validation accuracy vs k is shown in Figures 4.a and 4.b.

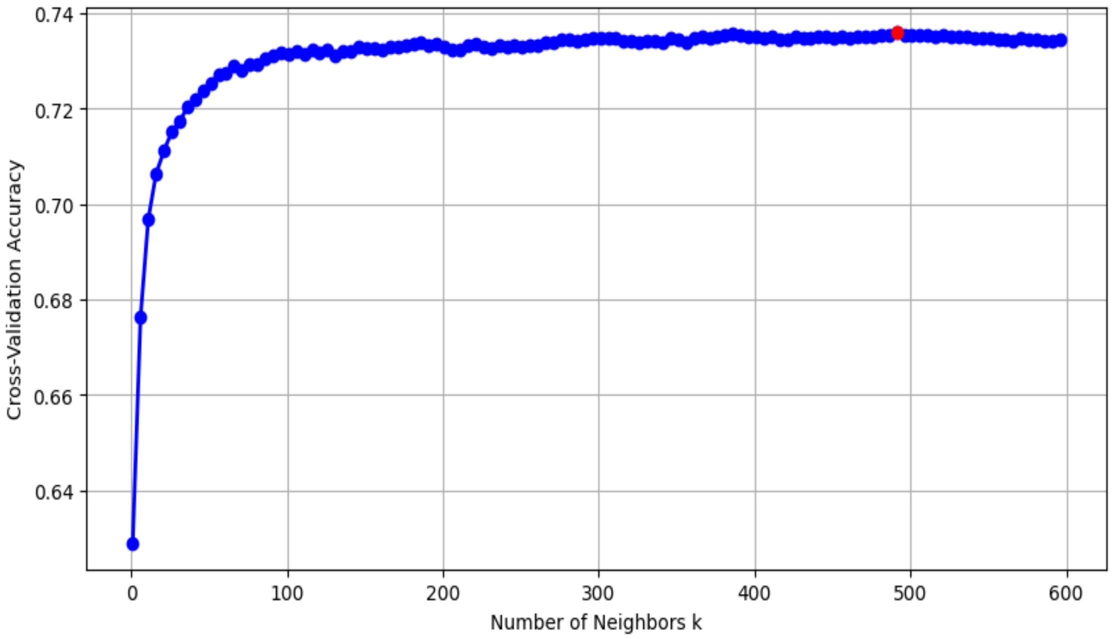

Figure 4.a: Cross-validation accuracy for KNN model with no PCA

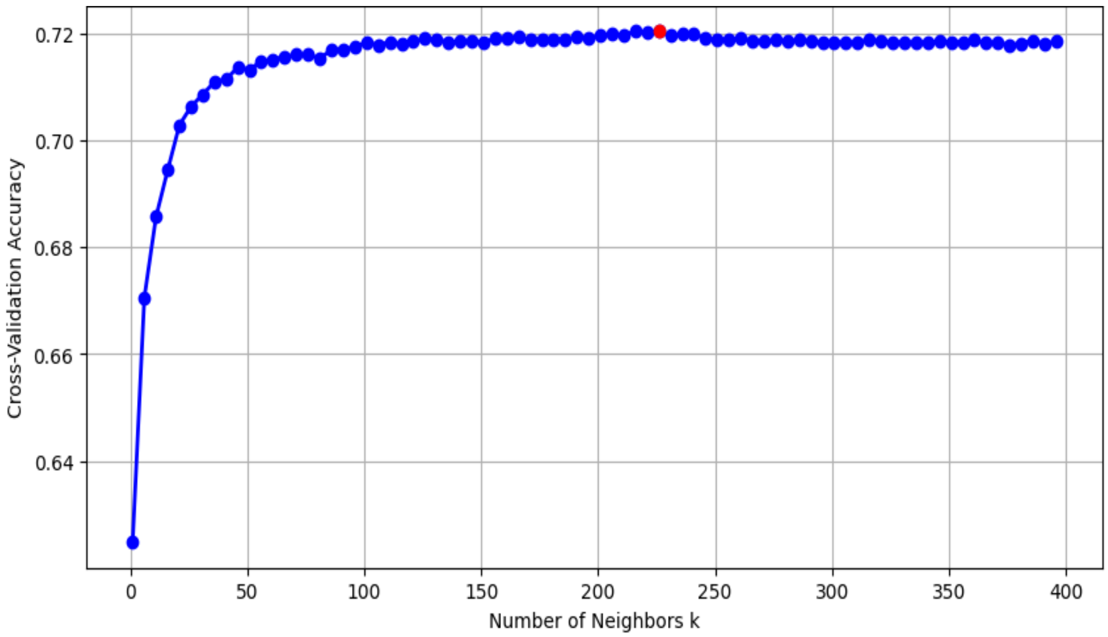

Figure 4.b: Cross-validation accuracy for KNN model with PCA

Random Forest Classifier

In the Random Forest classifier, the team varied several parameters to identify the best model configuration. These included:

- Number of estimators: The number of trees in the forest

- Maximum number of features per leaf: The maximum number of features considered for splitting a node
- Maximum depth: The maximum depths of the tree
- Minimum samples per leaf: The minimum number of samples required to be at a leaf node
- Minimum samples per split: The minimum number of samples required to split an internal node
- Bootstrap: Whether bootstrap samples are used in the forest

AdaBoost

The team opted to train an AdaBoost model rather than the XGBOOST model featured in the project's proposal. For the AdaBoost classifier, the base estimator used was decision stumps. The team varied the number of estimators and the learning rate through grid search cross validation. This approach allowed the team to fine-tune the model.

Multi-layer Perceptron (MLP) Neural Network

The MLP Neural Network required careful tuning of several parameters:

- Hidden layer sizes
- Activation functions
- Regularization parameters
- Learning rates
- Momentum

Early stopping was utilized during grid search cross-validation to prevent overfitting and reduce training time. This technique stops training when the mode's performance on a randomly selected portion of the training set starts to degrade.

2.1.6 Model Evaluation

The team determined that there are two main methods for how the models should be trained, and this is largely dependent upon the firm retention strategy. The first retention strategy dubbed the 'Rehiring Strategy,' is self-evident. Companies employing this strategy would utilize the model's prediction to plan for rehiring the position. A few benefits of this strategy, if effective, would be more accurate hiring and budget forecasts, and less time spent missing headcount which leads to reduced total overtime paid. If the model predicts multiple employees leaving the same position, or team, it could strengthen the forecasts and allow the company to prepare better. In this situation, the firm would like the model to focus on high accuracy and reducing the number of false positives (high precision). False positives would hinder the firm utilizing this strategy since the company would plan to hire a new employee, but the old employee was not considering leaving the firm.

The second retention strategy dubbed the 'Firefighter Strategy,' covers the opposite response to a positive model prediction. In this strategy, the firm would focus on incentivization for employees that they believe will be leaving the company. Studies show it is cheaper to retain employees than rehiring new employees. A benefit of this strategy, if effective, would be a reduction in hiring and onboarding costs. While the first strategy focuses on high accuracy and precision, this strategy focuses on high accuracy and reducing the number of false negatives (recall or sensitivity). False negatives would negatively affect firms employing this strategy, as they would not be able to attempt to convince the worker that will leave.

Depending on the firm's strategy, models should be selected based on these metrics. In general, the F-1 score represents a balance between precision and recall and can be proven as a basic number to select models. In an ideal state, the company would likely have multiple models. Companies feel differently toward different employees – a firm has superstar employees they would wish to retain, and underperformers who they would rather replace. By utilizing two models, focused on each strategy, the firm could evaluate the likelihood of employees staying or leaving for both models, and prefer the score for the model that fits their strategy tailored to that employee. This would be more feasible within smaller companies due to care and attention to employee, however large companies would also have an advantage with this strategy along with their advantage of having more employee data for training the model.

Due to this, accuracy, f-1, sensitivity, specificity, and precision scores were included in the model results for each model. Model evaluation results are shown for both PCA and non-PCA data below in Tables 1 & 2.

Table 1: Model Result without PCA

Model	**Accuracy**	**F-1**	**Precision**	**Recall**	**Specificity**
Logistic Regression	0.754	0.739	0.742	0.735	0.771
Gaussian Naïve-Bayes	0.730	0.730	**0.773**	0.692	0.772
KNN (k=491)	0.741	0.726	0.724	0.728	0.752
Ridge Classifier	0.754	0.739	0.740	0.738	0.768
Random Forest	0.748	**0.760**	0.766	0.755	0.741
Neural Network	0.750	0.744	0.719	**0.771**	0.731
AdaBoost	**0.764**	0.749	0.751	0.746	**0.779**
SVM with 'rbf' kernel	0.748	0.732	0.735	0.729	0.765

Table 2: Model Result with PCA (n=28)

Model	**Accuracy**	**F-1**	**Precision**	**Recall**	**Specificity**
Logistic Regression	0.581	0.558	0.555	0.561	0.598
Gaussian Naïve-Bayes	0.608	0.623	0.570	**0.687**	0.538
KNN (k=226)	0.597	0.574	0.572	0.577	0.615
Ridge Classifier	0.582	0.561	0.556	0.566	0.596
Random Forest	**0.640**	**0.611**	**0.624**	0.598	**0.678**
Neural Network	0.620	0.601	0.596	0.606	0.633
AdaBoost	0.621	0.602	0.596	0.608	0.597
SVM with 'rbf' kernel	0.626	0.609	0.600	0.618	0.633

Comparing Models

From the results, models trained on data without PCA performed far better than the models trained on PCA data. This is likely since the non-PCA data had 41 predictors compared to the 28 predictors in PCA. Typically, the goal of PCA is to reduce dimensionality by a large factor to save time and maintain generally high accuracy. Since the dimensionality wasn't reduced significantly, however, the total explained variance of PCA was 86%, the models trained on the full dataset were able to account for more variance within the data and thus resulted in more accurate models across the board. In general, PCA is still viable for model training and selection, especially if the firm has far more predictor variables than the datasets the team used from Kaggle.

Selecting the Best Model

To choose the best models, the team selected one single best model for accuracy and F-1 score, and a best model for each precision and recall. All models were chosen based off the scores without PCA due to superior results. From the results, the model with the highest accuracy was AdaBoost and the highest F-1 score was Random Forest. The Gaussian Naïve-Bayes model had the highest recall score, and the MLP Neural Network had the highest precision score. Interestingly, the leaders in each category were from different models, whereas in the PCA results, Random Forest dominated every category except recall, for which Gaussian Naïve-Bayes produced the best results.

Model Accuracy

In general, model accuracy was decent, but not great. A model which randomly assigns labels of 1 or 0 would achieve an accuracy of 0.5. The trained models did not extend past 0.764, which shows that the model is not incredibly accurate. This could be explained by multiple potential factors.

The first factor is that the data may have been synthesized in a random way, leading to difficulty in separation. The team could not determine the sourcing method for the dataset, so this point exists as a possibility.

The second factor is that there are multiple variables absent in the data that could better encapsulate the employee's motivation to stay or leave. As mentioned previously, there are many factors that contribute to an employee's motivation to leave their current employment. Not all these factors can be captured through demographic or survey data. PwC compiled a list of five predictors that make up a 'Resignation Equation' for employee turnover, these include job fulfillment, ability to be their true self, financial compensation, and level of care from their team and manager [4].

One example of how we could improve the data collection methods is from the ISYE 6740 Homework where the team utilized a Divorce Predictors data set to train a model [5]. This Divorce Predictors data featured survey responses from couples based on questions that represented the individual's feelings toward their partner across 54 well thought-out questions. Something similar could be implemented at a company through an incentivized Employee Opinion Survey (EOS), and responses could be gathered for specific usage in model training and prediction. The EOS questionnaire would need to be well designed to encapsulate the employee's motivation but could prove to be a better predictor of attrition than demographic data.

Another example of data that can be used is a graphical representation of an individual, such as who and/or what they associate with, various types of social media data. Utilizing graphical community-based data may give more insight into a person's personality, drivers, opinions, and response to stimuli. A final example of data that can be used is external factors such as economic conditions or industry-specific events which may affect attrition.

Finally, a third potential factor could be that the methods the team used during model training were not customized to the specific model uses, such as focusing on highest cross-validation score, or not focusing on tuning recall and precision individually.

3.1 Conclusion and Next Steps

Overall, the results were clear and helped lead the team to the next direction. It is evident that PCA is not necessary on the sample data unless more predictor variables are present. Furthermore, since each best score was dominated by different models, utilizing methods such as boosting or bagging between these various models can be leveraged to achieve even stronger results. More data collection is not strictly necessary, but having more angles of context may result in higher model accuracy.

Throughout model training and evaluation, the team faced many difficulties. Most notably, the team was significantly limited by time. Time is often the most limiting factor of the machine learning process. With more time available, larger grid searches could be conducted. Larger grid searches along with returning the scoring results of each parameter combination would be used to visualize the bias-variance tradeoff to choose more optimal parameters that aren’t overfit to the training data. Opting for models that are less overfit will result in higher accuracy to unseen data.

Additionally, once parameters were chosen, the team could evaluate model accuracy, f-1, precision, and recall by varying the cutoff threshold. In training and evaluation, the team used a default cutoff value of 0.5, however in an ideal state, full metric curves for each would be computed and factored into model selection.

Furthermore, with more time, the team would be able to tailor specific models to optimize for each precision and recall, rather than one model focused on general accuracy and f-1 score. These models would be specifically tailored to the firm’s needs.

There are multiple solutions for the time complexity issue that the team will consider in future machine learning projects. By utilizing more powerful computers or cloud computing clusters, the team would spend less downtime waiting for model training and evaluation. Another method is using weighted sampling to sample the training dataset and assign weights based on the prior distribution of the data. Weighted sampling would reduce both the time and energy required for the computations.

Overall, the results of the model training were promising. The team proved that machine learning models can predict attrition based on basic data. While the results were not ideal, the team gained experience in working with various machine learning models and gained ideas for how to make more effective prediction models by leveraging multiple techniques.

The results prove that companies that choose to develop a bespoke attrition prediction model would create a competitive advantage over rival companies that don’t. Additionally, these companies could gather data from within their company on routine bases and focus their machine learning models to be more accurate for their specific uses. Finally, depending on the model(s) the firm utilizes, companies can perform analyses on factors leading to attrition to support the creation robust retention strategies. While attrition is a growing problem in specific industries, attrition prediction models can help companies mitigate the negatives of turnover and create a competitive advantage.

4.1 References

[1] Monthly quite levels and rates by industry and region, Bureau of Labor Statistics, https://www.bls.gov/news.release/jolts.t04.htm.

[2] Benchmarking, Society of Human Resource Management, https://www.shrm.org/topics-tools/research/shrm-benchmarking#accordion-a5599cb1d9-item-b5dbc3c3b3.

[3] Employee Attrition Dataset, Kaggle, https://www.kaggle.com/datasets/stealthtechnologies/employee-attrition-dataset?select=train.csv.

[4] Global Workforce Hopes and Fears Survey 2022, PwC, Global Workforce Hopes and Fears Survey 2022 - Web report (pwc.com)

[5] Divorce Predictors Data Set, Divorce Predictors data set - UCI Machine Learning Repository