Noah Vriese
Springboard Data Science

Final Project Report:
Human Resources Attrition Prediction

Problem Statement

In the United States in 2020, the average annual employee turnover rate reached 57.3% [1]. This represents an increase by 14.7% from 42.6% just four years prior (2016). Combine this phenomenon with the expenses associated with costly employee replacement and training—6-9 months of an employee's salary on average—and the financial incentive for employee retention becomes apparent. At first glance, the issue of attrition seems personal to the employee and subject to a confounding number of external factors beyond the business's control. But what can be done to address this? Through leveraging the often-rich trove of internal company data, could we identify the most important factors influencing an employee's personal decision to leave a company? Could we also then identify those employees at the greatest risk of attrition and divert Human Resources (HR) assets to improve their outcomes?

Fortunately, International Business Machines - Canada (IBM) has published an HR database containing anonymized workplace demographic information comprising 24 initial features including salary, age, commute distance, department, attrition, etc. for roughly 4,000 employees. This data was leveraged to train a supervised machine learning model for the purpose of identifying the primary factors influencing employee attrition and predicting which employees are within the companies twentieth percentile of likelihood to turnover.

The final Random Forest Classifier (RFC) model developed during this exploration can use 20 features to achieve 95.2% accuracy and 97.5% recall when predicting an employee's likelihood of attrition within the next year. This model can be adapted and deployed for HR department use where sufficient employee demographic data is available.

Data Wrangling

The raw dataset from IBM (sourced from Kaggle [2]) contained 4,410 rows and 24 columns (features) which are described in Table 1 below. Four additional files contained employee timesheet data ("in_time.csv", "out_time.csv") and survey data from the employee and their manager respectively ("employee_survey_data.csv", "manager_survey_data.csv").

Data wrangling operations began with feature extraction from the four supplementary files to merge 9 additional features to the raw dataset (Table 1). Each employee's number of sick days, and mean, max, sum and standard deviation of daily hours worked were extracted from the supplementary timesheet data for the year covered. In addition, employee and manager survey data were joined to the raw dataset to complete the initial feature extraction.

Next, non-English units were converted for clarity (ex. kilometers to miles), and records containing missing or null values were dropped from the raw DataFrame provided they represented < 2% of the 4,410 rows present (< ~90 records). The final DataFrame shape was 4382 rows by 33 columns.

Table 1. Data Features, type, and brief description from the cleaned DataFrame.

Feature	Data Type	Description
Age	Integer	Employees current age
Attrition	Categorical: nominal	Attrition in the last year: (Y/N)
BusinessTravel	Categorical: nominal	Frequency of business travel
Department	Categorical: nominal	Department within company
DistanceFromHome	Float	Office distance from home
Education	Categorical: ordinal	Education level
EducationField	Categorical: nominal	Field of education
EmployeeCount	Integer	Employee count (constant)
Gender	Categorical: nominal	Employee gender
JobLevel	Categorical: ordinal	Job level at company: (1-5)
JobRole	Categorical: nominal	Job title
MaritalStatus	Categorical: nominal	Marital status of employee
MonthlyIncome	Float	Monthly income
NumCompaniesWorked	Integer	Total number of companies worked for
Over18	Categorical: nominal	Whether employee is over 18 years of age: (Y/N)
PercentSalaryHike	Float	Percent salary hike within last year
PerformanceRating	Categorical: ordinal	Performance rating of employee given by manager
StandardHours	Integer	Base number of hours per week for employee
StockOptionLevel	Categorical: ordinal	Stock option level for the employee
TotalWorkingYears	Integer	Total number of working years for the employee
TrainingTimesLastYear	Integer	Number of trainings attended by employee
YearsAtCompany	Integer	Total number of years at the current company
YearsSinceLastPromotion	Integer	Number of years since last promotion
YearsWithCurrManager	Integer	Number of years with current manager
MeanHrsWorked	Float	Mean number of hours worked per week
MaxHrsWorked	Float	Max number of hours worked per week
SumHrsWorked	Float	Total number of hours worked in the last year
StdHrsWorked	Float	Standard deviation of hours worked in the last year
SickDays	Integer	Number of sick days taken in the last year
EnvironmentSatisfaction	Categorical: ordinal	Employee surveyed environment satisfaction
JobSatisfaction	Categorical: ordinal	Employee surveyed job satisfaction
WorkLifeBalance	Categorical: ordinal	Employee surveyed work-life balance
JobInvolvement	Categorical: ordinal	Job involvement as rated by manager

Exploratory Data Analysis (EDA)

A thorough EDA of the cleaned DataFrame was conducted for the purpose of identifying relationships between features which could improve model performance.

The binary target variable for predictive modeling ('Attrition') was selected during EDA. Then percentage differences in independent variales (ex. 'Age', 'SickDays', 'WorkLifeBalance') between attritional and non-attritional employees were ranked and used as an initial measure of predictive power (Figure 1).

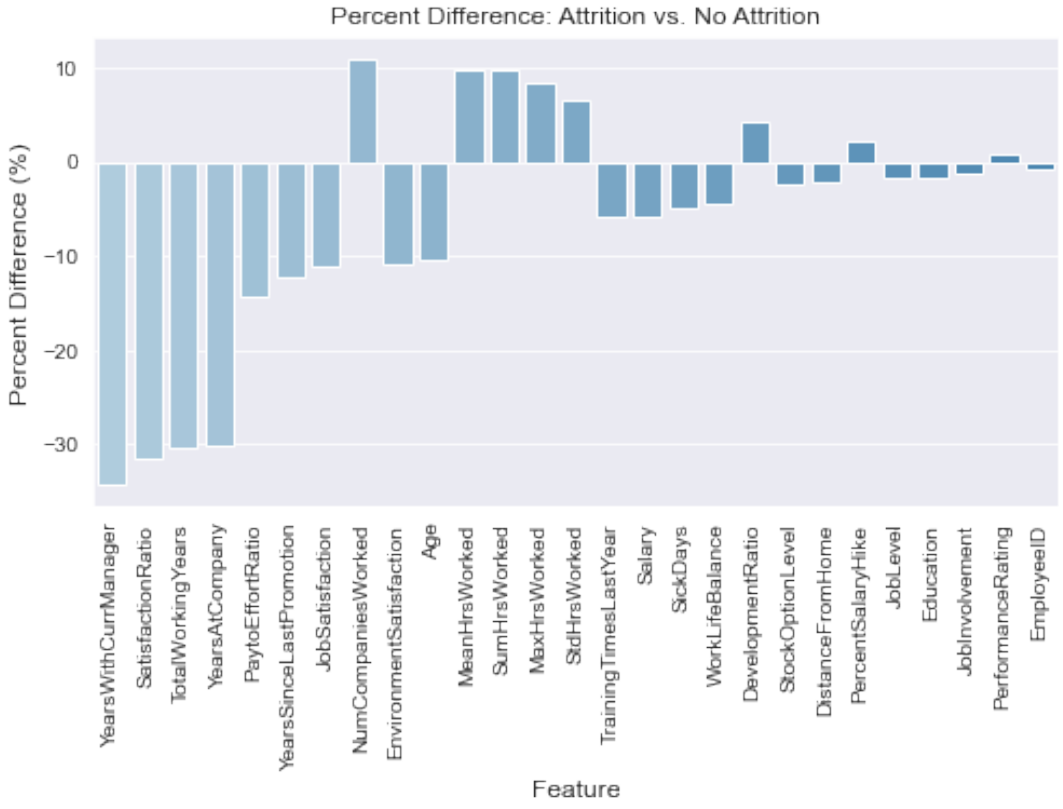

Figure 1. Percentage difference in supporting feature values between attritional and non-attritional employees.

Feature Engineering

Once an initial ranking of predictive power was established, higher ranked (more predictive) features were used to engineer new features. Four additional features were engineered: 'PaytoEffortRatio', 'SatisfactionRatio', 'DevelopmentRatio', and 'FlightRatio'. 'FlightRatio' is ranked most predictive and not shown on Figure 1 due to the +132.5% difference between attritional and non-attritional employees. Formulations for each engineered feature are shown below:

$$\text{PaytoEffortRatio} = \frac{\text{Salary}}{\text{SumHrsWorked}}$$

$$\text{SatisfactionRatio} = \frac{\text{WorkLifeBalance} * \text{EnvironmentSatisfaction} * \text{JobSatisfaction}}{\text{MeanHrsWorked}}$$

$$\text{DevelopmentRatio} = \frac{\text{TrainingTimesLastYear} + \text{JobLevel} + \text{StockOptionLevel}}{1 + \text{YearsSinceLastPromotion}}$$

$$\text{FlightRatio} = \frac{\text{NumCompaniesWorked}}{\text{TotalWorkingYears} * \text{EnvironmentSatisfaction}}$$

Because employee 'Age' ranks high (11th) among independent variables influencing attrition, ages were grouped into 7 bins of 5-year ranges and the percentage of attritional employees within each age bin was then calculated (Figure 2). Using this binning, a categorical 'AgeGroup' feature was added to the DataFrame.

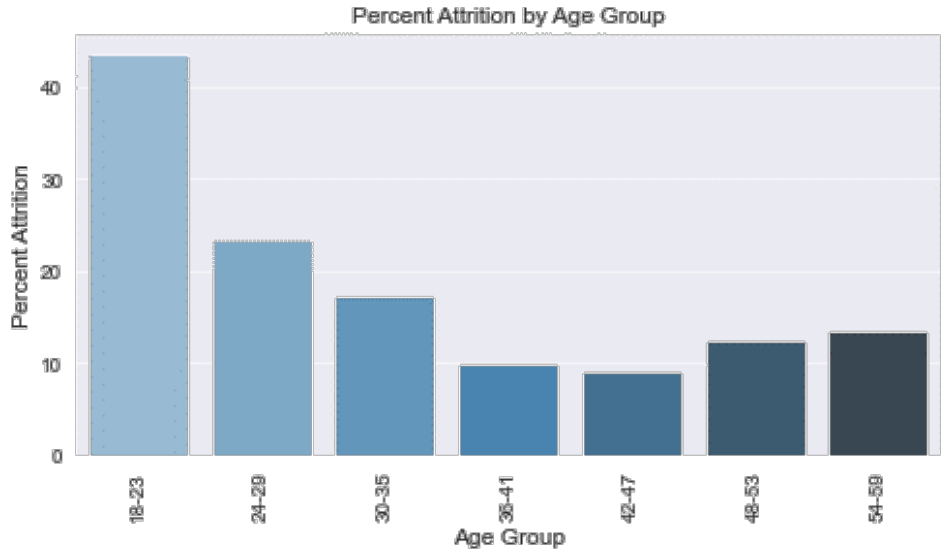

Figure 2. Percentage of attritional employees by age group.

Testing for Statistical Significance

Similarly, graphical comparison of the difference in the distribution of 'TotalWorkingYears' (ranked 4th), 'NumCompaniesWorked' (9th) and 'MeanHrsWorked' (12th) by attritional and non-attritional employees was continued with a series of boxplots (Figure 3). The differences observed between attritional and non-attritional employees were determined to be statistically significant (Table 2). Notably, employees who quit or left the company in the past year possessed on average fewer working years, fewer former employers, and worked more hours daily.

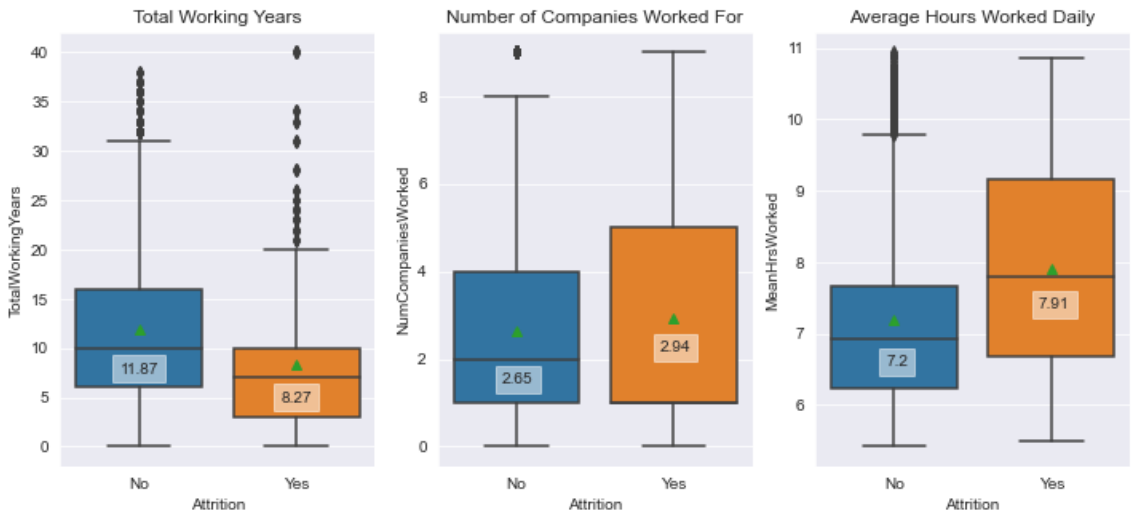

Figure 3. Box and whisker plot distributions (mean values labeled) for attritional and non-attritional employees by feature.

Table 2. Statistical measures for the difference in mean value between attritional and non-attritional employees.

Feature	p-value	t-value	Significant Difference?
TotalWorkingYears	1.165 E-29	-11.394	Y, $p << 0.05$
NumCompaniesWorked	4.572 E-3	2.837	Y, $p < 0.05$
MeanHrsWorked	3.287 E-39	13.230	Y, $p << 0.05$

Visual Exploration of Predictive Features

Numerical predictive features including 'MeanHrsWorked' and 'Salary' were plotted with hue indicating 'Attrition' and point size proportional to 'FlightRatio' (defined above) to qualitatively identify notable clusters. As shown in Figure 4 below, many attritional employees fall on the lower half of the salary distribution and are especially concentrated on the lower-right quadrant of the plot. This cluster is indicative of employees who are generally under-compensated and overworked. Notably, employees with the highest 'FlightRatio' fall within the lower-right quadrant.

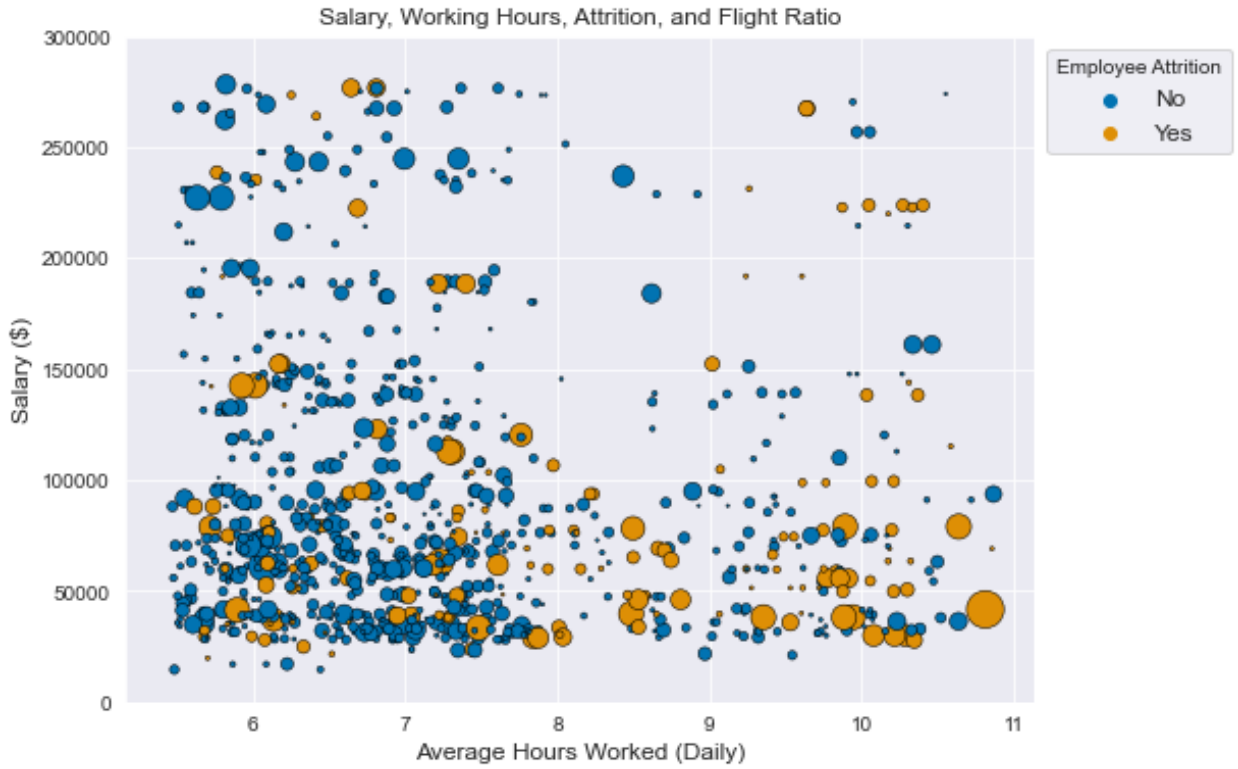

Figure 4. Scatterplot of 'MeanHrsWorked' (x-axis) vs. 'Salary' (y-axis). Hue indicative of employee attrition in the last year. Point size proportional to engineered feature 'FlightRatio'.

Predictive Model Development

Directly prior to model development, the HR data was preprocessed which included 70%-30% train-test splitting and stratification of the train-test split to ensure an approximately equal class balance of the target variable 'Attrition'. Categorical features were dummy encoded via Pandas 'get_dummies', and numerical features were scaled to unit variance using Scikit-Learn's 'StandardScaler'. The encoded train and test arrays were then exported for model development.

Baseline Models

Five standard classification algorithms were evaluated for relative performance during baseline model selection: Logistic Regression (LogReg), K-Nearest-Neighbors (KNN), Gradient Boosting Classifier (GB), Decision Tree, and Random Forest Classifier (RFC). The 5-fold Cross-Validation (CV) score was calculated on the training split for each out-of-the-box model (no hyperparameter/parameter tuning) and compared with the other baseline models for a relative performance score (Figure 5). RFC yielded the highest CV score, indicative of superior generalization without significant overfitting, and was therefore selected for further development. The baseline RFC without hyperparameter tuning was designated 'Model A' for later performance evaluation.

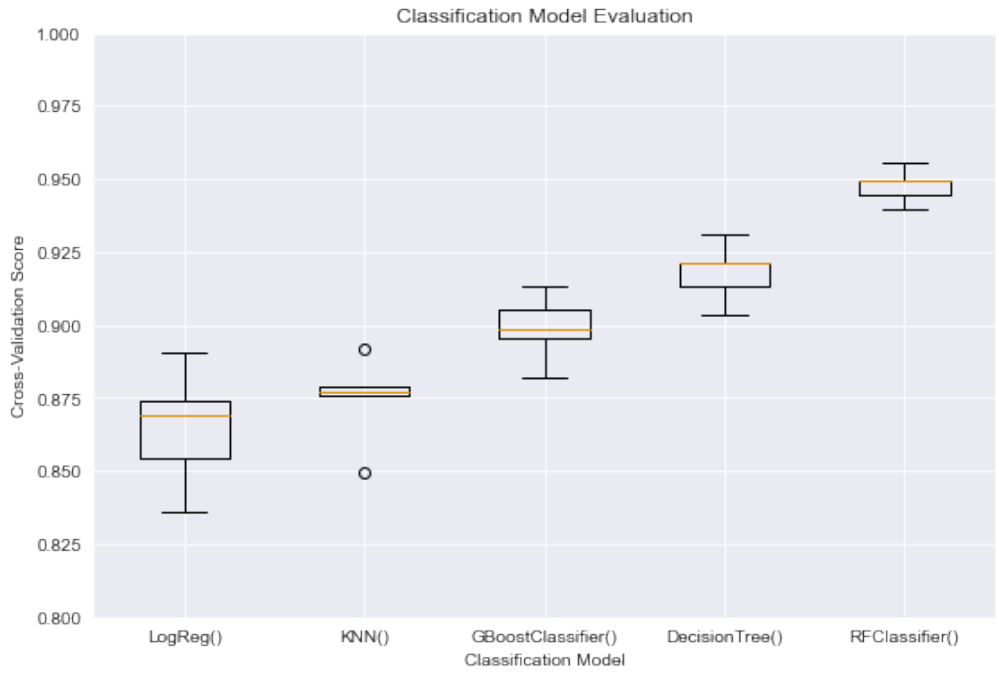

Figure 5. Cross-validation scores (5-fold) calculated for five baseline classification algorithms on the train split. RFC (CV-score: 0.948) performs best among the baseline model candidates.

Hyperparameter Tuning

A randomized grid search was implemented to improve the performance of Model A (RFC - baseline) through hyperparameter tuning to build Model B (RFC - tuned). The value ranges considered, and optimized values determined through 5-fold CV using Scikit-Learn's 'RandomizedSearchCV' are displayed in Table 3. In total, 500 model fits were evaluated including 5 folds for each of the 100 model candidates.

Table 3. RandomizedSearchCV hyperparameter ranges considered and optimized values implemented for the tuned RFC model ('Model B').

Hyperparameter	**Description**	**Value Range:** (optimized value)
n_estimators	Total number of trees	[200, 400, 600, 800, 1000]: (200)
max_features	Number of features to consider at each split	['auto', 'sqrt']: ('auto')
max_depth	Max number of levels in tree	[20, 65, 110, 155, 200]: (110)
min_samples_split	Min number of samples to split a node	[2, 5, 7]: (2)
min_samples_leaf	Min number of samples at each leaf node	[1, 2, 4]: (1)

Feature Selection

After encoding, the HR DataFrame included 74 features, which could be reduced through feature selection to reduce model dimensionality and improve processing time. As the best-performing model (Model B – RFC-tuned) was a RandomForestClassifier, RFC feature importance ranking was selected as a method to eliminate features offering little predictive power.

Therefore, once its performance was scored and determined to be reasonably high, Model B (RFC - tuned), was used to rank feature importance in predicting the target variable 'Attrition' (Figure 6).

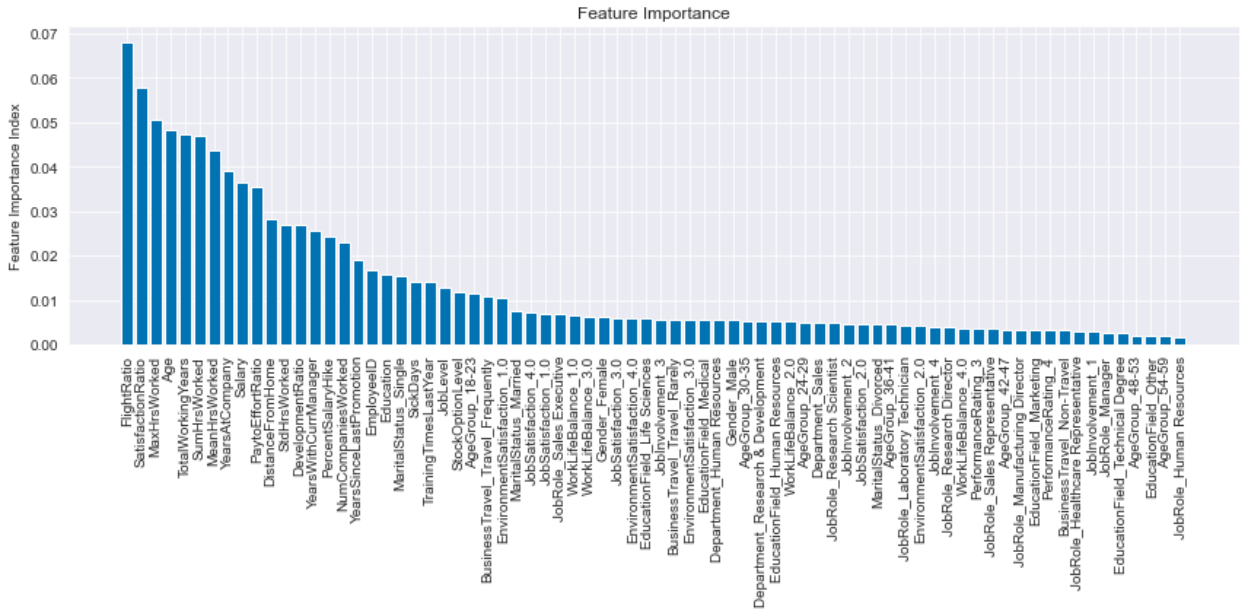

Figure 6. Feature importance rankings influencing employee attrition as calculated by the tuned RFC model.

Model Evaluation

In total, 5 models were evaluated for relative performance via the following metrics: accuracy, precision, recall, F1-score, ROC AUC, and confusion matrix (Figure 8). A summary of the tested models, and their performance metrics are listed in Table 4, and displayed graphically in Figure 7. Note that Models 'A', 'B' and 'D' used 74 features, while Models 'C' and 'E' were trained using only the 20 highest-ranked features by RFC feature importance (Figure 6).

Model 'C' was determined to be the best overall model for deployment due to similar or better performance relative to the 74-feature models with significantly less dimensionality.

Table 4. Evaluation metrics for models considered. Bolded/italicized values indicate best-in-class.

Metric	**Model A** RFC – base	**Model B** RFC – tuned	**Model C** RFC – tuned – 20 best features	**Model D** LogReg – base	**Model E** LogReg – tuned – 20 best features
Accuracy	0.961	***0.964***	0.952	0.856	0.856
Precision	***0.982***	0.977	0.975	0.591	0.646
Recall	0.774	***0.797***	0.722	0.354	0.241
F1-Score	0.865	***0.878***	0.829	0.442	0.351
ROC AUC	***0.989***	***0.989***	0.980	0.801	0.765

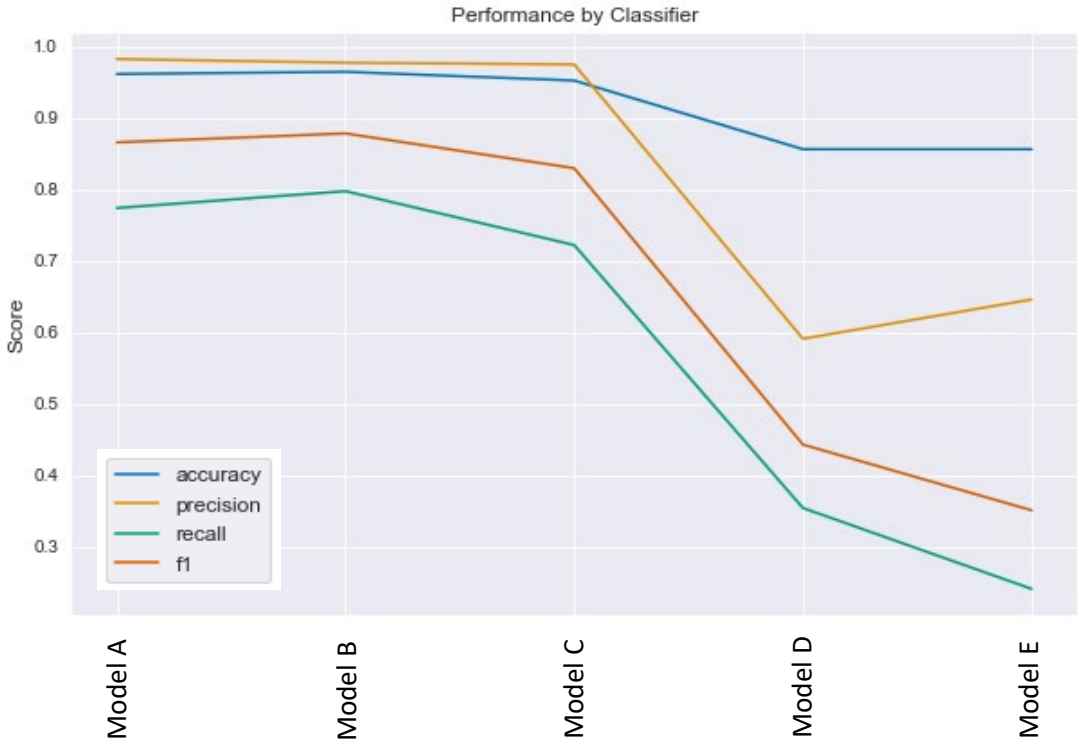

Figure 7. Performance by evaluation metric for the 5 classification models tested.

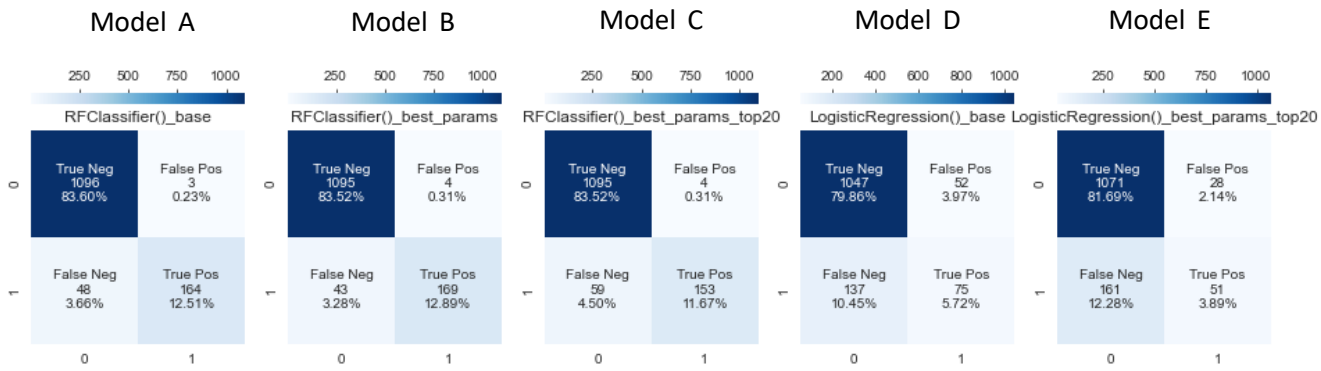

Figure 8. Confusion matrices for the 5 classification models evaluated.

Conclusions

Without significant hyperparameter optimization, the Random Forest Classifier model performs best among the 5 baseline classification algorithms tested (Logistic Regression, K-Nearest-Neighbors, Gradient Boosting Classifier, Decision Tree, Random Forest Classifier). Notably, RFC performance was improved only marginally with hyperparameter optimization (see Table 4: Model 'A' vs. Model 'B').

Arguably, the greatest improvement in model design was in feature reduction. Model 'C' utilized only 20 of the original 74 predictors and scored similarly to the 74 feature RFC models (Models 'A' and 'B') within the performance metrics considered. Therefore Model 'C' was selected as the 'best' overall model for implementation and was used to answer the two questions proposed for this work.

Responses to Questions Posed in the Problem Statement

I. What are the 5 most important features influencing the probability of employee attrition?

Based upon the feature importance ranking generated by the optimized RFC model, the five most predictive features when determining the probability of employee attrition are:

1. 'FlightRatio': a measure of an employee's number of companies worked for - per year
2. 'SatisfactionRatio': a measure of an employee's work-life balance and job satisfaction
3. 'MaxHrsWorked': the maximum number of hours worked by an employee in a given week
4. 'Age': Age of the employee.
5. 'TotalWorkingYears': Total number of years the employee has held any job

Note that two of these features ('FlightRatio', 'SatisfactionRatio') were engineered, see *Feature Engineering* section for full descriptions of each.

II. Can we identify employees within the upper 20th percentile who are most likely to leave the company?

Using the selected 'best' overall model (Model 'C'), attrition probability scores were calculated for all employees. Employees falling within the upper 20th percentile of attrition probability were flagged to be the focus of HR resources. The company-wide distribution of attrition probabilities is shown in Figure 9.

Note that the range considered with the 20th percentile includes employees with relatively low attrition probabilities (as low as ~0.25). The threshold for HR consideration could therefore be considered for adjustment to optimize resources.

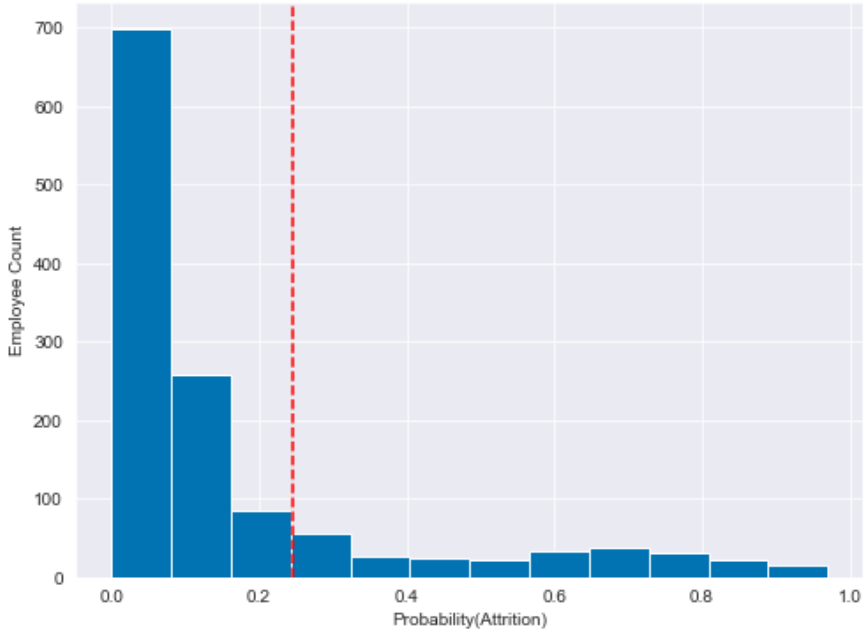

Figure 9. Employee attrition probability histogram as predicted by Model 'B'. Employees to the right of the vertical red line fall within the top 20th percentile of attrition probability.

Future Research

The high evaluation scores reflected in Models 'A', 'B', and 'C' suggest that when applied to common HR department data, machine learning can be an effective solution to HR resource management, and potentially employee attrition. While the model proposed for deployment does not in itself improve employee outcomes, it does aid HR representatives in the task of actively monitoring employee satisfaction and therefore is likely to improve employee retention when used as a predictive tool.

As noted above, the threshold of employee attrition probability should be a topic of future research for optimization, likely on a company-by-company basis to avoid flagging employees at a low probability of attrition. Instead of using a percentile threshold, HR departments could tailor their efforts to a select number of employees (ex. top 100 most attritional) depending upon department capacity.

Additionally, this model is well suited for deployment as a real-time monitoring or dashboard tool for HR departments, depending upon the availability of internal employee data. A lighter version of this model could also be developed for companies which possess less available employee demographic data.

References

[1] - Ariella, S. (2022). (rep.). *27 Us Employee Turnover Statistics [2022]: Average Employee Turnover Rate, Industry Comparisons, And Trends*. Zippia. Retrieved September 1, 2022, from https://www.zippia.com/advice/employee-turnover-statistics/#:~:text=The%20employee%20turnover%20rate%20in,2016%20to%2057.3%25%20in%202020.

[2] – Choudhary, V. (2018, August 10). *HR analytics case study*. Kaggle. Retrieved September 5, 2022, from https://www.kaggle.com/datasets/vjchoudhary7/hr-analytics-case-study