

Predicting Employee Attrition

LANCE ENSMINGER
2019

Table of Contents

Introduction	3
Objectives	5
Data	7
Methodology	19
Results	31
Analysis	34
Deliverables	43
References	45
Appendix	47

Introduction

Introduction - Attrition


- Attrition cost employers an estimated \$600 billion in 2018
- In the form of
 - Lost productivity
 - Replacement costs
 - Temporary employment
 - Training costs
- Per departing employee, this cost is about 1/5th of the employee's salary
- Some of the reasons that employees leave
 - Lack of career development
 - Desire for better work-life balance
 - Negative manager behavior
 - Personal well-being
 - Compensation and/or benefits
- Find a method to predict employee attrition

Objectives

Objectives

- Primary
 - Identify and train a classification model to predict employee attrition
- Secondary
 - Provide actionable insight
 - Variable importance
 - Estimate marginal probability after changing variables' values

Data



Environment

Interpretation

Data
Preparation

Environment

- Simulated by IBM data scientists
- Data associated with employees to encourage research about who leaves a company and who stays
- The format of the data is a comma separated values file
 - Easily readable into R
- There are 1,470 observations and 35 variables, 1 of which is an identifier

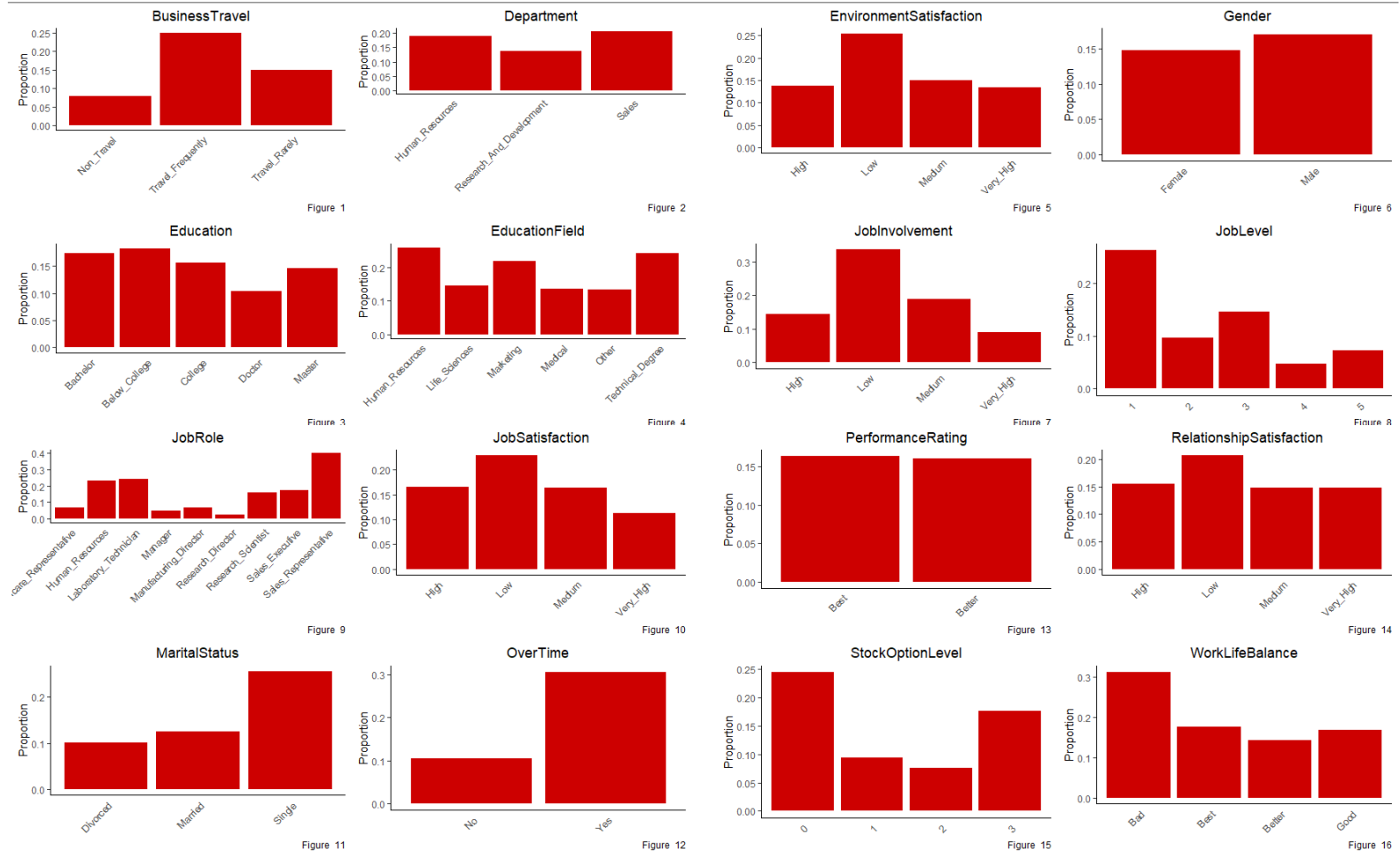
Interpretation

- There were no missing values
- 3 variables had zero variance and were removed
 - EmployeeCount
 - Over18
 - StandardHours
- Categorical variables
 - Bar charts of the proportion of those that left vs those that did not, by category
- Numerical variables
 - Histograms
 - Violin charts
 - Correlation plots

Interpretation – Categorical Variables – Bar Charts

- Bar charts in Figures 1-16 are used to understand the variables with categories that had little variation or significant variation among categories
 - This was only used to gain a better understanding of the data
 - No decisions were made to remove variables based on this visualization since it does not consider variable interactions
- Observations
 - Most variables had at least 1 category with significant variation
 - Variables with little variance among categories
 - Gender
 - PerformanceRating
 - Categories with the highest proportion that left the company
 - JobRole of SalesRepresentatives (0.40)
 - Accepted OverTime (0.31)
 - Perceived a Bad WorkLifeBalance (0.31)

Interpretation – Categorical Variables



Interpretation – Numerical Variables – Histograms

- Histograms in Figures 17-28 were used to determine if there were any numerical variables that were skewed, had little variance, or if anything looked odd
- Observations
 - Most numerical variables were skewed and would need to be transformed, if methods that work better with normally distributed variables were used, then centered and scaled to standardize the data

Interpretation – Numerical Variables – Histograms

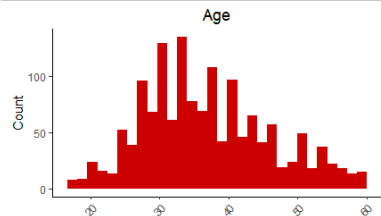


Figure 17

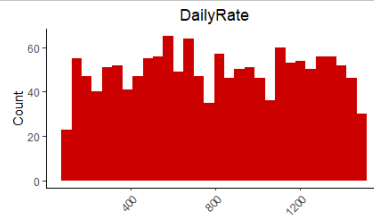


Figure 18

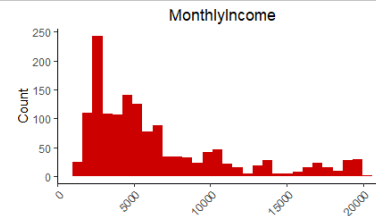


Figure 21

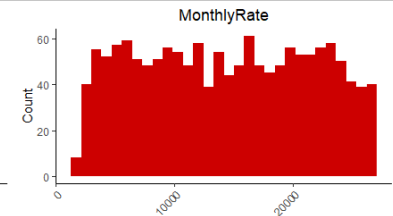


Figure 22

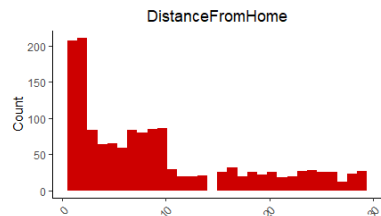


Figure 19

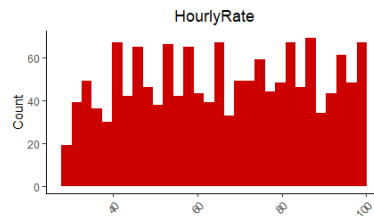


Figure 20

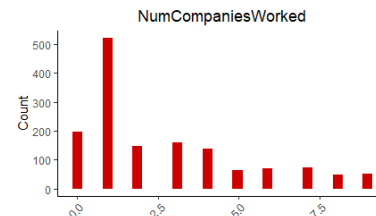


Figure 23

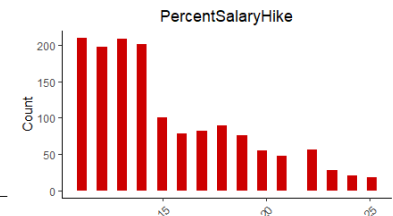


Figure 24

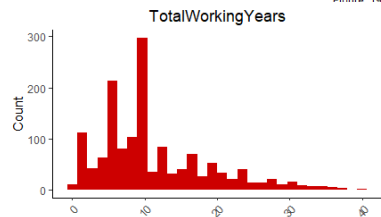


Figure 25

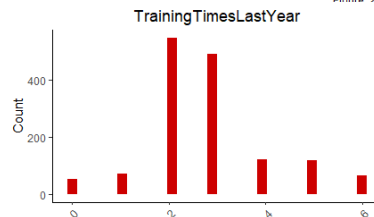


Figure 26

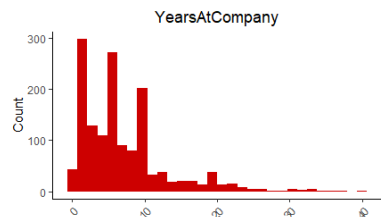


Figure 27

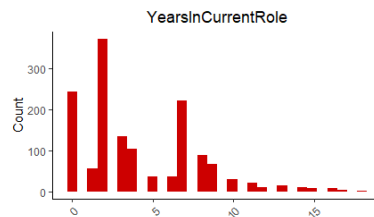


Figure 28

Interpretation – Numerical Variables – Violin Charts

- Violin charts in Figures 29-40 were used to glean anything about those who left the company versus those who did not
- Observations
 - Tendencies of those who left
 - Younger
 - Work farther from home
 - Lower MonthlyIncome
 - Worked for more companies
 - Fewer TotalWorkingYears
 - Fewer YearsAtCompany
 - Fewer YearsInCurrentRole

Interpretation – Numerical Variables – Violin Charts

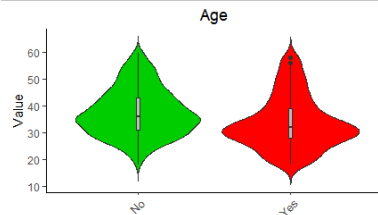


Figure 29

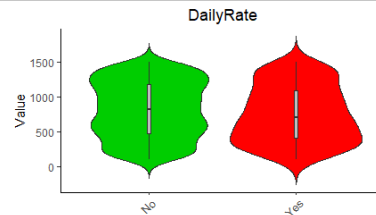


Figure 30

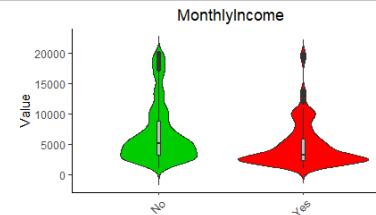


Figure 33

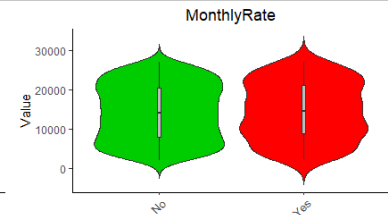


Figure 34

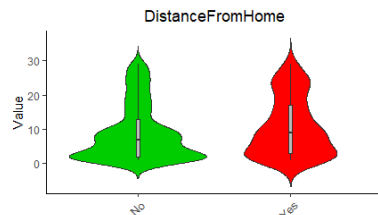


Figure 31

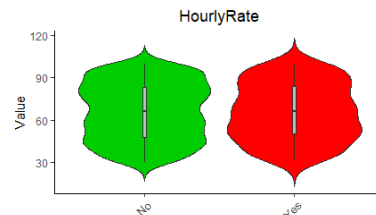


Figure 32

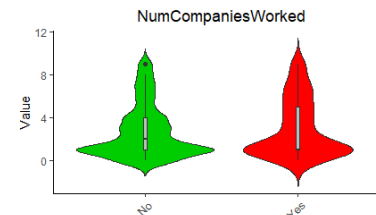


Figure 35

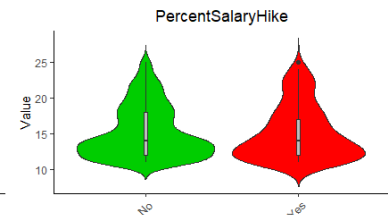


Figure 36



Figure 37

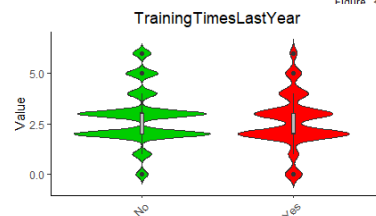


Figure 38

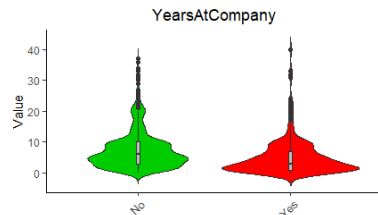


Figure 39

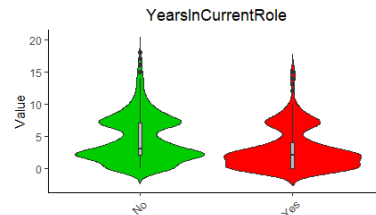
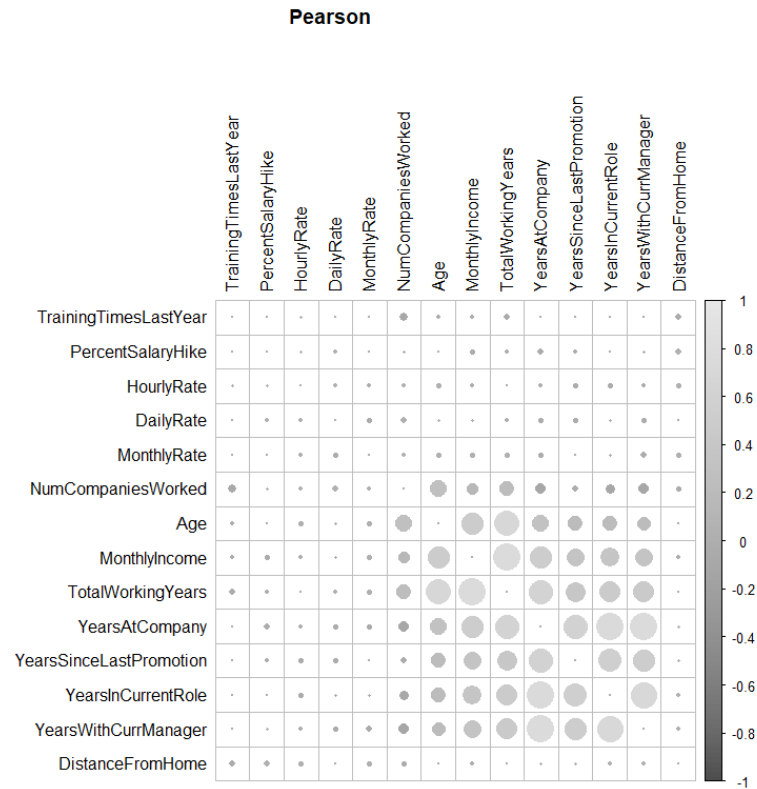


Figure 40

Interpretation – Numerical Variables – Correlation

- Correlations can indicate features that are possible candidates for removal
- Since there were some variables that had skewness, I looked at Pearson, Kendall, and Spearman correlations
 - Showing only the Pearson correlation plot in Figure 41, since the results were similar
- Observations
 - Variables with higher correlations
 - YearsWithCurrManager
 - YearsSinceLastPromotion
 - YearsInCurrentRole
 - YearsAtCompany
 - TotalWorkingYears
 - MonthlyIncome
 - Age
 - NumCompaniesWorkedFor
- Further analysis
 - Variables with correlation > 0.7
 - TotalWorkingYears
 - YearsInCurrentRole
 - YearsWithCurrManager
 - These variables were identified as possible candidates for removal

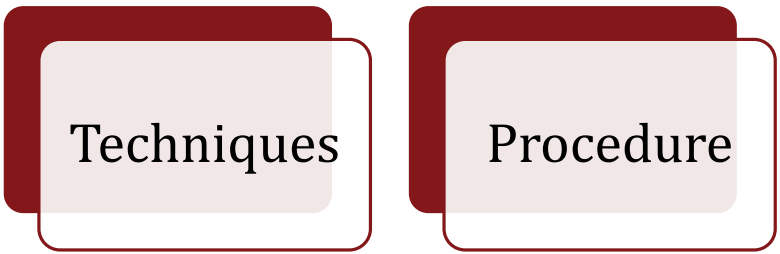
Interpretation – Numerical Variables – Correlation



Preprocessing

- All categorical variables were given dummy variables
- Since several variables were skewed, BoxCox transformations were applied to the variables, the transformed variables were then centered and scaled
- Split the data into train and test data
 - Randomly select 75% of the data for training
 - Use the remaining 25% for testing
- Use k-fold cross-validation with n repeats for hyperparameter tuning and model training
 - $k = 10$
 - $n = 5$

Methodology



Techniques

The diagram consists of two light pink rectangular boxes with rounded corners, each containing text. These boxes are positioned side-by-side and are partially overlaid by a dark red rectangular shape that is wider than the boxes and has rounded corners. This dark red shape is positioned behind the pink boxes, creating a layered effect. The entire diagram is set against a white background.

Procedure

Techniques

- Methods
 - Logistic Regression
 - Random Forests
 - Boosted Trees
 - Support Vector Machines
 - Linear
 - Radial
- Loss metric for model training
 - Log-loss

Techniques

- Diagnostic metrics for evaluating model performance on test data
 - Log-loss
 - Accuracy
 - Kappa
 - Sensitivity
 - Specificity
 - Precision
 - Recall
 - F1 Score

Techniques – Logistic Regression

- Regression model
 - $\ln \frac{\pi}{1-\pi} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$
- Probability transformation
 - $\pi = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n}}$
- Outcome classified as 1 when, $\pi \geq$ some classification threshold, otherwise it is classified as 0
- Penalized logistic regression
 - Used to combat overfitting
 - Ridge
 - Penalty: $\lambda \sum_{j=1}^p \beta_j^2$
 - Regularizes the impact of features that could cause overfitting and minimizes the effect of colinear features
 - LASSO
 - Penalty: $\lambda \sum_{j=1}^p |\beta_j|$
 - May eliminate some features that are correlated with others by setting their coefficients to 0
 - Combination
 - Penalty: $\lambda \left[(1 - \alpha) \frac{1}{2} \sum_{j=1}^p \beta_j^2 + \alpha \sum_{j=1}^p |\beta_j| \right]$
 - Applied penalized logistic regression using the Combination
 - Tuning hyperparameters are λ and α

Techniques – Logistic Regression

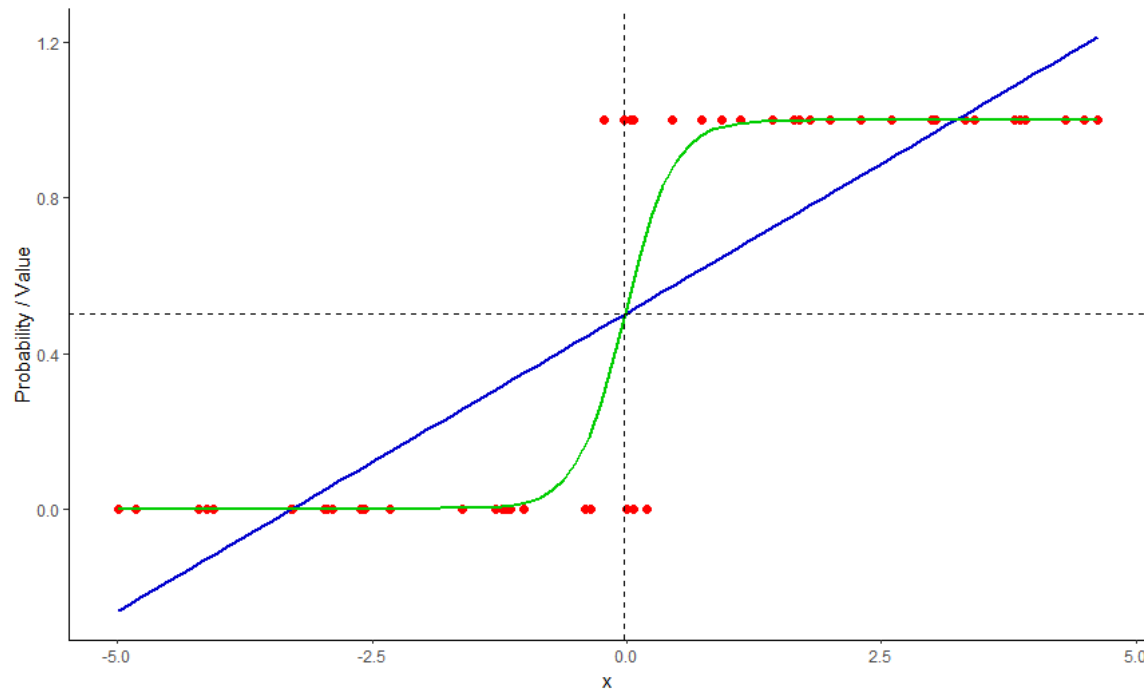


Figure 42

- Blue line: Ordinary Least Squares Regression
- Green line: Logistic Regression
- Each classification threshold set to 0.5
- Similar results when outcome variable is symmetric
- Less error using logistic regression
- Logistic regression can also provide a probability that an occurrence is 1 or 0

Techniques – Logistic Regression

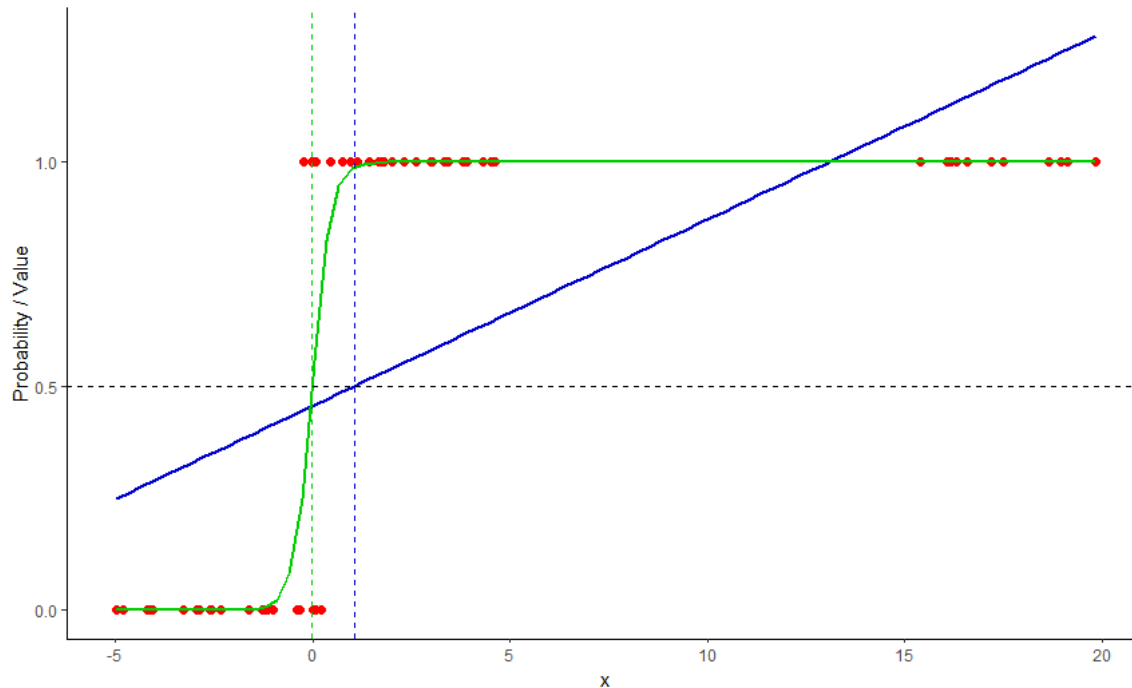


Figure 43

- Blue line: Ordinary Least Squares Regression
- Green line: Logistic Regression
- Each classification threshold set to 0.5
- Dissimilar results when outcome variable is asymmetric

Techniques – Random Forests

- Forest of decision trees
- Each tree is built using randomly selected occurrences and randomly selected variables from the original set
- Individual trees are built using a measure of impurity to decide the order of the nodes
 - Gini index
 - $I_{Gini}(C) = 1 - \sum_{i=1}^C (p_i)^2$, where p_i = probability of class i
 - May be weighted if the decision has a different number of total outcomes
- Tuning hyperparameter is the number of randomly selected features available to be chosen at each split
- Each occurrence to be predicted is calculated on each tree in the forest
 - The outcome that occurs most frequently in the forest is the predicted outcome

Techniques – Gradient Boosting

- Ensemble of trees
- Each tree's residuals are based on the prior tree's predicted probabilities
- The final prediction is the sum of the weighted outcomes of all the trees
- The original tree is a leaf representing an average of the observed values
 - For classification this is $\ln \frac{\pi}{1-\pi}$, where $\pi = P(Y = 1)$
 - Each leaf's outcome is also $\ln \frac{\pi}{1-\pi}$
- To help prevent overfitting, a learning rate is tuned to scale each tree
- Other tuning hyperparameters
 - Number of trees
 - Number of leaves
 - Maximum depth
 - Minimum weight of child
 - Number of available features
 - Number of observations

Techniques – Support Vector Machines

- Find hyperplanes that best separate the classes
 - The best hyperplanes are those that have the greatest margin between the hyperplane and the closest observation to the hyperplane
- Linear
 - Has 1 hyperparameter to tune
 - The weight of misclassification, which helps to avoid overfitting by letting some misclassifications occur creating a larger margin hyperplane
- Radial
 - Has 2 hyperparameters to tune
 - The first is like the weight of misclassification used in the Linear version
 - The second defines how much influence any particular training sample has on the hyperplane
 - High values place more value on the samples closest to the hyperplane
 - This could cause overfitting
 - Low values place more value on the samples farthest from the hyperplane
 - This could cause a more linear hyperplane and not allow the hyperplane to adjust to the complexity of data

Techniques – Loss Function for Tuning Hyperparameters During Model Training

- Log-loss
- A diagnostic metric that represents the overall fit of the model to the data
- Used when the predicted function is a sigmoid function
 - Since the sigmoid function is not linear and the transformation of the probability being modeled is either $-\infty$ or $+\infty$, the typical methods, like mean squared error are not meaningful
- $\text{Log-loss} = -y \times \ln(\hat{y}) - (1 - y) \times \ln(1 - \hat{y})$
- Advantage of Log-loss over other diagnostic metrics
 - Ability to consider how close, or far away, the prediction is from the observation by utilizing the probability instead of just considering the binary outcome

Techniques – Testing – Other Diagnostic Metrics

- Accuracy =
$$\frac{\text{true positives} + \text{true negatives}}{\text{true positives} + \text{true negatives} + \text{false positives} + \text{false negatives}}$$
- Kappa =
$$\frac{\text{probability of observed agreement} + \text{probability of expected agreement}}{1 - \text{probability of expected agreement}}$$
- Sensitivity =
$$\frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$
- Specificity =
$$\frac{\text{true negatives}}{\text{true negatives} + \text{false positives}}$$
- Precision =
$$\frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$
- Recall =
$$\frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$
- F1 Score =
$$\frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Procedure

- Applied these techniques using R
 - Primarily using Caret
- Create cross-validated sample sets
- Train the models
- Predict the outcomes using the best tuned models on the test set, setting the classification threshold to the probability that maximizes both Sensitivity and Specificity
- Analyze the predictions from the test set

Results



Training

The diagram consists of two identical rectangular blocks positioned side-by-side. Each block has a dark red background with a lighter red rounded rectangle centered on it. The word 'Training' is written in black serif font inside the left block, and 'Testing' is written in black serif font inside the right block. A thin dark red border surrounds the lighter red area of each block.

Testing

Training

Table 1

	Logistic Regression	Random Forest	Gradient Boosted Trees	Support Vector Machines - Linear	Support Vector Machines - Radial
Training Parameters	alpha = 0.1 lambda = 0.0208	mtry = 40	nround = 150 max_depth = 1 eta = 0.3 gamma = 0 colsample_bytree = 0.8 min_child_weight = 1 subsample = 0.75	C = 1	C = 1 Sigma = 0.0069
CV Log-loss	0.2926	0.3476	0.2927	0.2992	0.3023

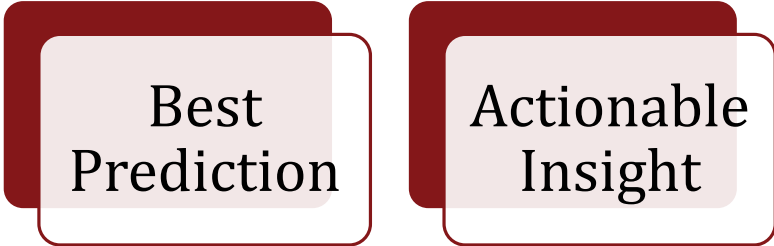
- After removing the 3 variables that had the highest Pearson's correlation, TotalWorkingYears, YearsInCurrentRole, and YearsWithCurrentManager, the cross-validated log-loss values for all methods were greater than in Table 1
 - Hence, these 3 variables remain in the data sets

Testing

Table 2

	Logistic Regression	Random Forest	Gradient Boosted Trees	Support Vector Machines - Linear	Support Vector Machines - Radial
Log Loss	0.3437	0.3737	0.3535	0.3550	0.3238
Accuracy	0.7989	0.8043	0.8125	0.8152	0.7989
Kappa	0.3824	0.3295	0.3926	0.4050	0.3963
Sensitivity	0.6491	0.5088	0.6140	0.6316	0.6842
Specificity	0.8264	0.8585	0.8489	0.8489	0.8199
Area Under the ROC Curve	0.8052	0.7477	0.7892	0.7892	0.8260
Optimal Cutoff	0.1948	0.2523	0.2108	0.2108	0.1740
Precision	0.4066	0.3973	0.4268	0.4337	0.4105
Recall	0.6491	0.5088	0.6140	0.6316	0.6842
Area under the Precision/Recall Curve	0.5582	0.4039	0.5518	0.5379	0.5700
F1 Score	0.5000	0.4462	0.5036	0.5143	0.5132

Analysis



Best
Prediction

Actionable
Insight

Best Prediction

- Comparing the cross-validated log-loss from the training data in Table 1, logistic regression and gradient boosted trees performed best, but both Support Vector Machine versions were not far behind
- Comparing the results from the testing data in Table 2, Support Vector Machine – Radial performed well in several categories, including log-loss
 - In addition to log-loss, it performs best by the Sensitivity metric
 - Sensitivity is important here, since we are interested in predicting those employees at risk of leaving and having a better sensitivity metric means that it is better at classifying the Yes Attrition (or 1) class, more true positives in relation to all positives in the observations
 - Specificity is lower, but this may be a sacrifice worth making if the real concern is those at risk of leaving

Table 3

		Glmnet		Random Forests		Gradient Boosted Trees		Support Vector Machine - Linear		Support Vector Machine - Radial	
Observations		Predictions		Predictions		Predictions		Predictions		Predictions	
		No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
	No	257	54	267	44	264	47	264	47	255	56
	Yes	20	37	28	29	22	35	21	36	18	39

- Table 3 shows the confusion matrix from the results of the test data

Actionable Insight

- One of the important aspects of this project is not only to be able to predict the classification, but also to provide some actionable insight regarding the variables and how they impact the probability of an employee leaving
 - Important variables
 - Use trained logistic regression model to provide a marginal probability estimate after changing the value of the inputs

Actionable Insight – Important Variables

Table 4

	Logistic Regression	Random Forest	Boosted Trees	Support Vector Machine - Linear	Support Vector Machine - Radial
Rank	Variable	Variable	Variable	Variable	Variable
1	JobInvolvement.Low	MonthlyIncome	MonthlyIncome	OverTime.Yes	OverTime.Yes
2	OverTime.No	Age	StockOptionLevel.0	OverTime.No	OverTime.No
3	OverTime.Yes	DailyRate	OverTime.Yes	MonthlyIncome	MonthlyIncome
4	JobRole.Manager	DistanceFromHome	OverTime.No	YearsAtCompany	YearsAtCompany
5	WorkLifeBalance.Bad	MonthlyRate	DailyRate	TotalWorkingYears	TotalWorkingYears

- In Table 4, I have only shown the top 5 important variables in each method for brevity
- With this truncated set, in each of the above variables you can see that some of the important variables are the same and some are not
- There are 3 important variables that are in the top 20 of all models, OverTime.Yes, OverTime.No, and StockOptionLevel.0
- So, just with this information we can say that if you want to increase the probability that an employee stays, reduce overtime and give them stock options

Actionable Insight – Marginal Probability Using Logistic Regression

Table 5

Rank	Variable	Coefficient	Rank	Variable	Coefficient
1	JobInvolvement.Low	0.9345	6	RelationshipSatisfaction.Low	0.6678
2	OverTime.No	-0.9087	7	EnvironmentSatisfaction.Low	0.6522
3	OverTime.Yes	0.9050	8	BusinessTravel.Travel_Frequently	0.6325
4	JobRole.Manager	-0.7962	9	EducationField.Technical_Degree	0.5920
5	WorkLifeBalance.Bad	0.7573	10	EducationField.Human_Resources	0.5744

- Table 5 shows the coefficients for the first 10 variables; the rest can be seen in Table A.1 in the appendix
- Use the logistic regression model to understand how a predicted outcome will change if an input's value is changed
- Logistic regression model
 - $\ln \frac{\pi}{1-\pi} = -2.1607 + 0.9345x_1 - 0.9087x_2 + 0.9050x_3 - 0.7962x_4 + 0.7573x_5 \dots + \beta_n x_n$

Actionable Insight – Marginal Probability Using Logistic Regression

- Be careful when interpreting this model
- We are interested in understanding the predicted probability
- The coefficients of this model are used to predict the natural logarithm of the odds
- To understand how the probability changes we must transform the natural logarithm of the odds to get the probability of each case and take the difference

Actionable Insight – Marginal Probability Using Logistic Regression

Table 6

Variable	EmployeeNumber = 1	Variable	EmployeeNumber = 1
BusinessTravel	Travel_Rarely	WorkLifeBalance	Bad
Department	Sales	Age	41
Education	College	DailyRate	1102
EducationField	Life_Sciences	DistanceFromHome	1
EnvironmentSatisfaction	Medium	HourlyRate	94
Gender	Female	MonthlyIncome	5993
JobInvolvement	High	MonthlyRate	19479
JobLevel	2	NumCompaniesWorked	8
JobRole	Sales_Executive	PercentSalaryHike	11
JobSatisfaction	Very_High	TotalWorkingYears	8
MaritalStatus	Single	TrainingTimesLastYear	0
OverTime	Yes	YearsAtCompany	6
PerformanceRating	Better	YearsInCurrentRole	4
RelationshipSatisfaction	Low	YearsSinceLastPromotion	0
StockOptionLevel	0	YearsWithCurrManager	5

- Table 6 shows a sample employee and their data that will be used to explain how to calculate the change in probability

Actionable Insight – Marginal Probability Using Logistic Regression

- Transform, center, and scale the numerical variables
- Make dummy variables for the categorical variables
- Use best tuned logistic regression model to find $\ln \frac{\pi}{1-\pi}$
- See appendix for dummy variables and adjusted data
 - The dummy variables, transformations, centering, and scaling have been added to Table A.2 in the appendix
 - Table A.3 in the appendix shows the λ , mean, and standard deviation used to make the adjustments

Actionable Insight – Marginal Probability Using Logistic Regression

- Applying this formula

$$\ln \frac{\pi}{1-\pi} = -2.1607 + 0.9345x_1 - 0.9087x_2 + 0.9050x_3 - 0.7962x_4 + 0.7573x_5 \dots + \beta_n x_n$$

to EmployeeNumber 1, $\ln \frac{\pi}{1-\pi} = 0.5756$

- Transforming this results in a probability that EmployeeNumber 1 leaves of $\pi = 0.6401$
- If we change OverTime.Yes to 0 and OverTime.No to 1, then $\ln \frac{\pi}{1-\pi} = -1.2381$ and $\pi = 0.2248$
- Holding all else constant, this is a vast improvement in probability (the probability decreases) that EmployeeNumber 1 will leave
- This can be done with more variables, or different variables providing the employer a tool that will help them keep the employees they find valuable

Deliverables

Deliverables

- An employer can now understand what factors are important
- How the logistic regression model can be used to quantify the impact on the probability of an employee leaving after making changes to the values of inputs
- This will help them increase retention of employees they want to keep
- This will also help them save a significant amount of money by not having to replace those employees

References

References

- Retention Report (2018). Retrieved from <http://info.workinstitute.com/2018retentionreport>
- Boushey, H. & Glynn, S. J. (2012). There Are Significant Business Costs to Replacing Employees. Retrieved from <https://www.americanprogress.org/issues/economy/reports/2012/11/16/44464/there-are-significant-business-costs-to-replacing-employees/>
- Kuhn, M. (2019). The caret Package. Retrieved from <https://topepo.github.io/caret/>
- Starmer, J. (2018). StatQuest with Josh Starmer. Retrieved from <https://www.youtube.com/user/joshstarmer>
- Winston, P. (2010). Artificial Intelligence, Lecture 16: Support Vector Machines. Retrieved from https://www.youtube.com/watch?v=_PwhiWxHK8o
- Kuhn, M. & Johnson, K. (2013). Applied Predictive Modeling. New York: Springer Science+Business Media
- RBF SVM parameters. Retrieved from https://scikit-learn.org/stable/auto_examples/svm/plot_rbf_parameters.html

Appendix

Logistic Regression Model – Coefficients

Table A.1

Variable	Coefficient	Variable	Coefficient
(Intercept)	-2.1607	JobRole.Research_Director	-0.2329
JobInvolvement.Low	0.9345	Age	-0.2160
OverTime.No	-0.9087	Department.Sales	0.2018
OverTime.Yes	0.9050	MaritalStatus.Divorced	-0.1908
JobRole.Manager	-0.7962	JobLevel.4	-0.1878
WorkLifeBalance.Bad	0.7573	YearsInCurrentRole	-0.1820
RelationshipSatisfaction.Low	0.6678	Education.Below_College	-0.1738
EnvironmentSatisfaction.Low	0.6522	TrainingTimesLastYear	-0.1598
BusinessTravel.Travel_Frequently	0.6325	Education.College	0.1592
EducationField.Technical_Degree	0.5920	DailyRate	-0.1568
EducationField.Human_Resources	0.5744	JobRole.Healthcare_Representative	-0.1525
JobSatisfaction.Low	0.5520	Gender.Female	-0.1378
BusinessTravel.Non_Travel	-0.5470	Gender.Male	0.1374
StockOptionLevel.0	0.5406	EducationField.Medical	-0.1321
JobRole.Research_Scientist	-0.4895	JobInvolvement.High	-0.1206
JobRole.Laboratory_Technician	0.4810	TotalWorkingYears	-0.1147
JobLevel.2	-0.4624	JobInvolvement.Medium	0.1104
JobSatisfaction.Very_High	-0.4621	EnvironmentSatisfaction.High	-0.1074
WorkLifeBalance.Better	-0.4069	WorkLifeBalance.Good	0.1072
Education.Doctor	-0.4027	EnvironmentSatisfaction.Very_High	-0.1017
JobRole.Sales_Executive	0.3948	JobSatisfaction.High	0.0964
JobInvolvement.Very_High	-0.3798	JobLevel.3	0.0813
MaritalStatus.Single	0.3703	JobSatisfaction.Medium	-0.0575
JobLevel.1	0.3488	RelationshipSatisfaction.Very_High	-0.0423
MonthlyIncome	-0.3472	RelationshipSatisfaction.High	-0.0349
DistanceFromHome	0.3416	BusinessTravel.Travel_Rarely	-0.0233
NumCompaniesWorked	0.3093	PercentSalaryHike	-0.0169
YearsSinceLastPromotion	0.3086	EducationField.Other	-0.0146
StockOptionLevel.2	-0.2901	MonthlyRate	0.0063
StockOptionLevel.1	-0.2740	EducationField.Life_Sciences	-0.0049
YearsWithCurrManager	-0.2607	JobLevel.5	0.0004
Department.Research_And_Development	-0.2440	JobRole.Human_Resources	0.0001
EducationField.Marketing	0.2367		

Adjusted Data for Sample

Table A.2

Variable	EmployeeNumber = 1	Variable	EmployeeNumber = 1
JobInvolvement.Low	0	TrainingTimesLastYear	-2.1712
OverTime.No	0	Education.College	1
OverTime.Yes	1	DailyRate	0.7609
JobRole.Manager	0	JobRole.Healthcare_Representative	0
WorkLifeBalance.Bad	1	Gender.Female	1
RelationshipSatisfaction.Low	1	Gender.Male	0
EnvironmentSatisfaction.Low	0	EducationField.Medical	0
BusinessTravel.Travel_Frequently	0	JobInvolvement.High	1
EducationField.Technical_Degree	0	TotalWorkingYears	-0.4215
EducationField.Human_Resources	0	JobInvolvement.Medium	0
JobSatisfaction.Low	0	EnvironmentSatisfaction.High	0
BusinessTravel.Non_Travel	0	WorkLifeBalance.Good	0
StockOptionLevel.0	1	EnvironmentSatisfaction.Very_High	0
JobRole.Research_Scientist	0	JobSatisfaction.High	0
JobRole.Laboratory_Technician	0	JobLevel.3	0
JobLevel.2	1	JobSatisfaction.Medium	0
JobSatisfaction.Very_High	1	RelationshipSatisfaction.Very_High	0
WorkLifeBalance.Better	0	RelationshipSatisfaction.High	0
Education.Doctor	0	BusinessTravel.Travel_Rarely	1
JobRole.Sales_Executive	1	PercentSalaryHike	-1.4974
JobInvolvement.Very_High	0	EducationField.Other	0
MaritalStatus.Single	1	MonthlyRate	0.7482
JobLevel.1	0	EducationField.Life_Sciences	1
MonthlyIncome	0.2848	JobLevel.5	0
DistanceFromHome	-1.5747	JobRole.Human_Resources	0
NumCompaniesWorked	2.1244	Department.Human_Resources	0
YearsSinceLastPromotion	-0.6789	Education.Bachelor	0
StockOptionLevel.2	0	Education.Master	0
StockOptionLevel.1	0	EnvironmentSatisfaction.Medium	1
YearsWithCurrManager	0.2458	JobRole.Manufacturing_Director	0
Department.Research_And_Development	0	JobRole.Sales_Representative	0
EducationField.Marketing	0	MaritalStatus.Married	0
JobRole.Research_Director	0	PerformanceRating.Better	1
Age	0.5253	PerformanceRating.Best	0
Department.Sales	1	RelationshipSatisfaction.Medium	0
MaritalStatus.Divorced	0	StockOptionLevel.3	0
JobLevel.4	0	WorkLifeBalance.Best	0
YearsInCurrentRole	-0.0633	HourlyRate	1.3520
Education.Below_College	0	YearsAtCompany	-0.1646

Transformed, Centered, and Scaled Data for Sample

Table A.3

Variable	Original Value	λ	Mean	Standard Deviation	Transformed Value	Transformed & Centered Value	Transformed, Centered, & Scaled Value
Age	41	0.2	5.240	0.511	5.508	0.268	0.525
DailyRate	1102	0.7	147.983	56.626	191.071	43.088	0.761
DistanceFromHome	1	0.1	1.971	1.251	0.000	-1.971	-1.575
HourlyRate	94	0.8	34.109	8.877	46.110	12.001	1.352
MonthlyIncome	5993	-0.2	4.088	0.119	4.122	0.034	0.285
MonthlyRate	19479	0.7	1122.668	418.861	1436.041	313.373	0.748
NumCompaniesWorked	8		2.693	2.498	8.000	5.307	2.124
PercentSalaryHike	11	-1.3	0.745	0.007	0.735	-0.010	-1.497
TotalWorkingYears	8		11.280	7.781	8.000	-3.280	-0.421
TrainingTimesLastYear	0		2.799	1.289	0.000	-2.799	-2.171
YearsAtCompany	6		7.008	6.127	6.000	-1.008	-0.165
YearsInCurrentRole	4		4.229	3.623	4.000	-0.229	-0.063
YearsSinceLastPromotion	0		2.188	3.222	0.000	-2.188	-0.679
YearsWithCurrManager	5		4.123	3.568	5.000	0.877	0.246