Case Study 2: Predicting Prospective Employee Salary and Attrition

Jeff Nguyen

06 December 2019

Introduction

DDSAnalytics has chosen multiple machine learning techniques to identify prospective candidate attrition and salary trends. Doing so will allow us to better select the right job seeker for our clients. Logistic and Naive Bayes classification models will be compared and one will be selected to predict candidate attrition. A successful classifier will have at least a 60% sensitivity and 60% specificity. A salary model using linear regression will also be used to identify salary trends given the features provided where the RMSE must below $3000. RStudio 1.2.1335 and R 3.6.1 will be used to manipulate, model, and present the data to the client.

Data Sources

Data for used for modeling training will be from `CaseStudy2-data.csv`, where prediction for `Attrition` will be performed on `CaseStudy2CompSet No Attrition.csv` and prediction on `MonthlyIncome` trends will be performed on `CaseStudyData2CompSet No Salary.csv`. Features that are numeric and discrete or character based will be converted to factors for ease of model ingestion.

Exploratory Data Analysis

Outlier and Duplicate Check

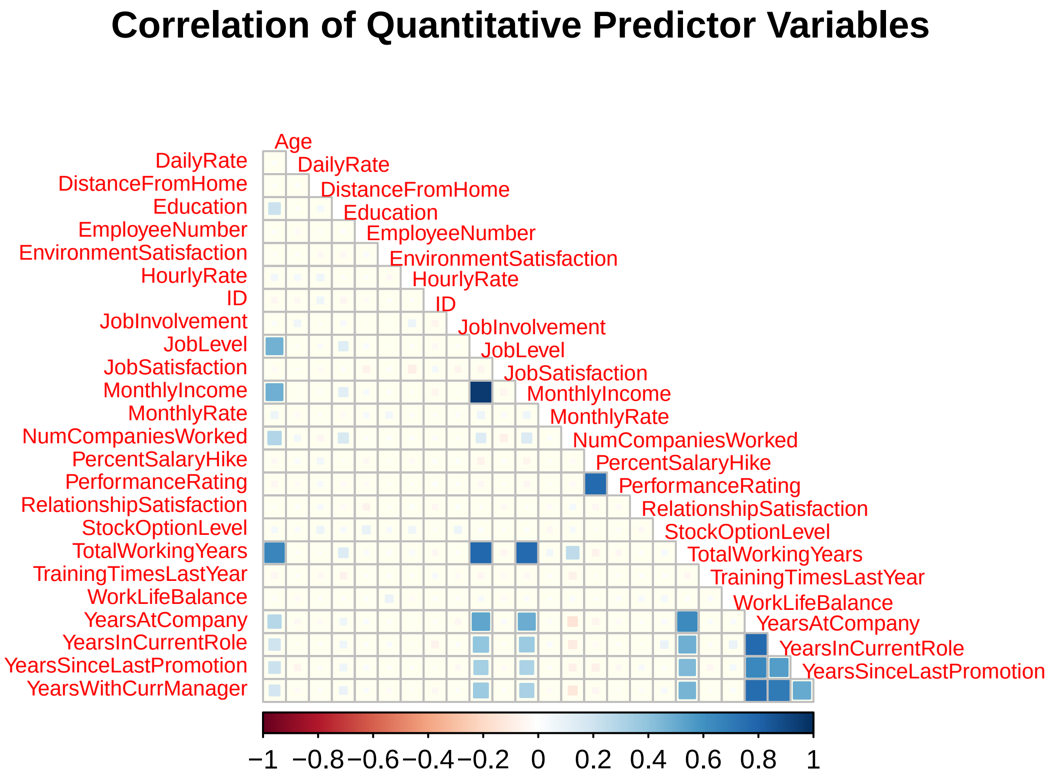
No duplicates or NAs were identified in the data provided. `get_dupes()` from the `janitor` package and `skim()` from the `skimr` package were used to identify duplicates, NA's, categorical levels, and quantitative distributions. Duplicates can inflate the values a model generates and it is best to remove these values before preceding to EDA or modeling.

Initial Variable Elimination

Based on exploratory data analysis `StandardHours`, `EmployeeCount`, and `Over18` have only one value. These features are removed from the data as they do not provide useful information to the models.

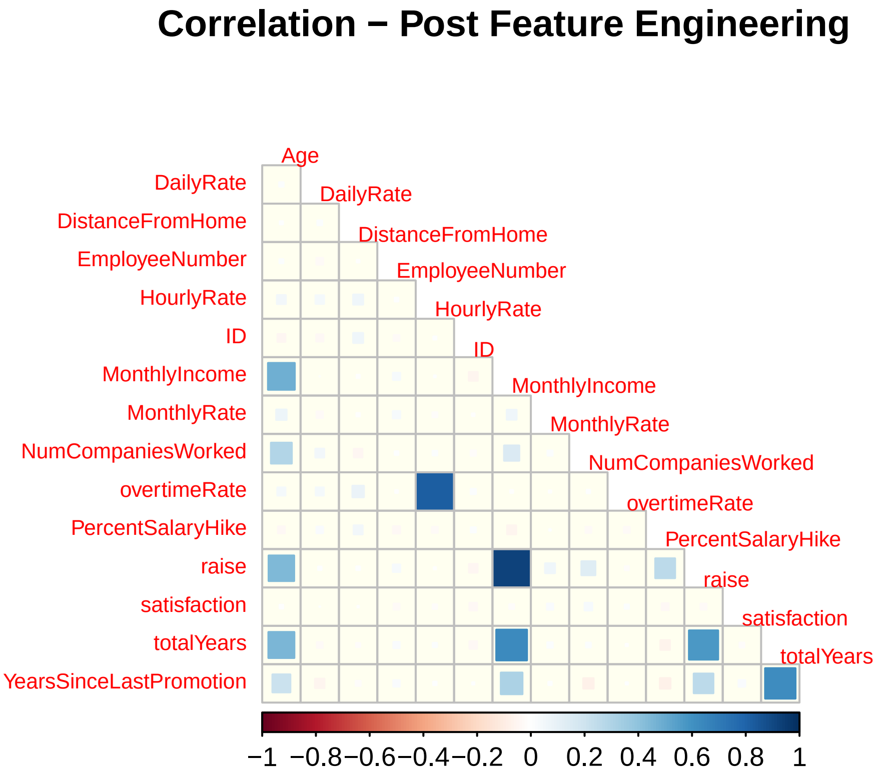
Correlation

Linear, Logistic, and Naive Bayes models require features that independent of each other. Correlation can be used to identify quantitative features that may display dependence by looking for features that are highly correlated. This is key if a model needs to establish causation, however for the purposes of prediction multicollinearity is more tolerable. This is the case because multicollinearity inflates the standard error values. When making predictions these values are less important as we are more focused on prediction. If needed, features that are collinear with each other can either be removed or combined to reduce or eliminate multicollinearity.



Correlation - Post Feature Engineering and Variable Removal

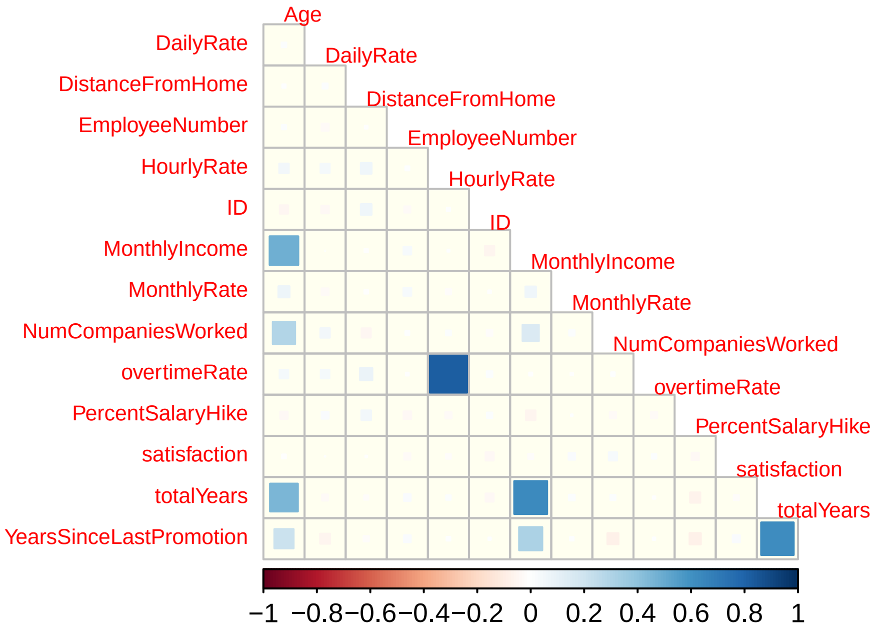
TotalWorkingYears, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager, and Age are highly correlated. The first four features involving time were combined and averaged out to create totalYears. This new variable intent is to reduce collinearity and provide a summarized term that captures an individual's time features. overtimeRate was created by identifying candidates that had the OverTime flag, then multiplying their HourlyRate was by 1.5; if the individual did not have the flag then their overtimeRate remained the same as their HourlyRate. The raise feature was created to capture the potential increase in an individual's MonthlyRate by multiplying their MontlyRate by PercentSalaryHike and dividing by 100. The last feature created is satisfaction- this feature is the average of RelationshipSatisfaction and EnvironmentSatisfaction and was created to capture the averaged satisfaction of a candidate.



Final Correlation - Variable Removal

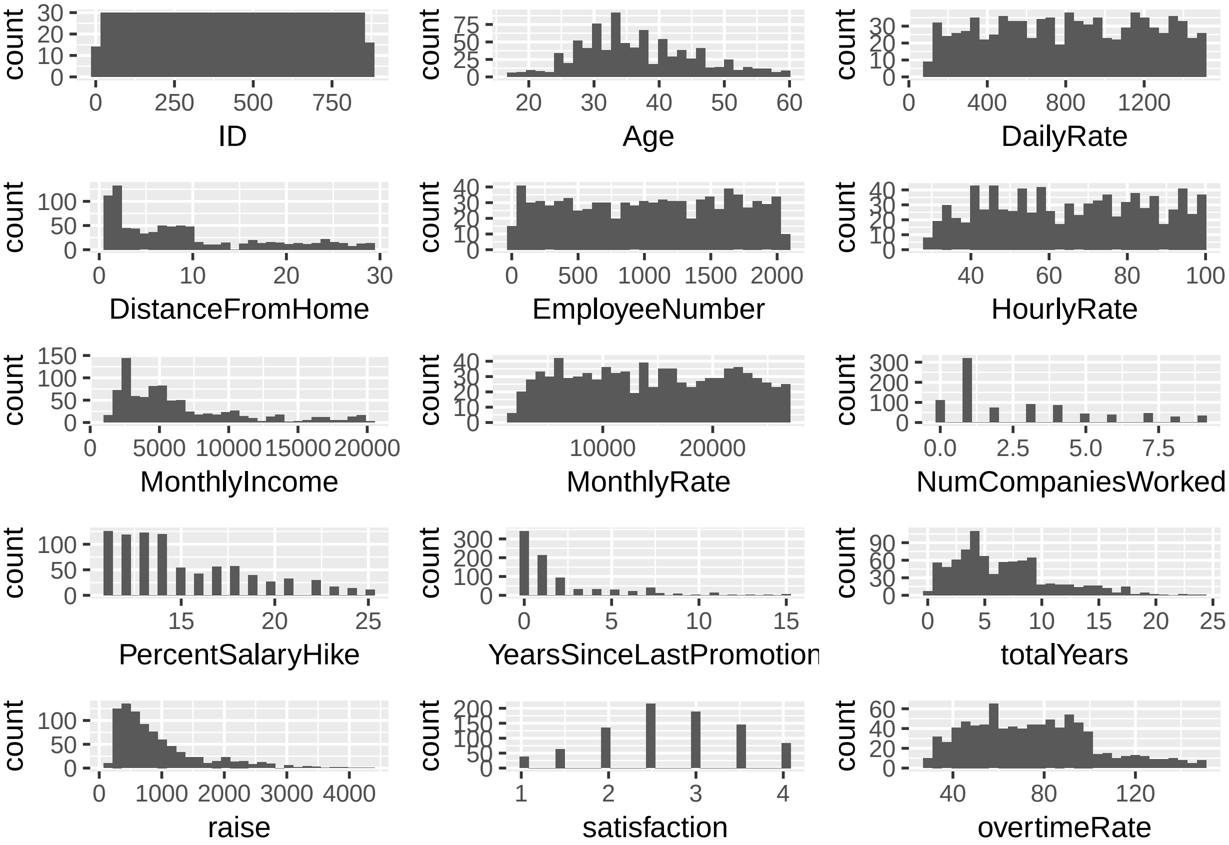
The previous correlation plot revealed that the newly created aggreagate features are highly correlated with their components and with `age`. The aggregation feature of `totalYears` and `raise` were also found to be highly collinear with several features; These features will be kept since they are not necessarily collinear with the response. Because we are preforming prediction, multicollinearity is less of a concern because we are not as interested in the standard errors a coefficients produce, however we are more interested in the predictions the models will generate. Since `raise` is highly colinear with the `MonthlyIncome` it will be removed from the salary prediction model since it is redundant.

Final Correlation of Quantitative Predictor Variables



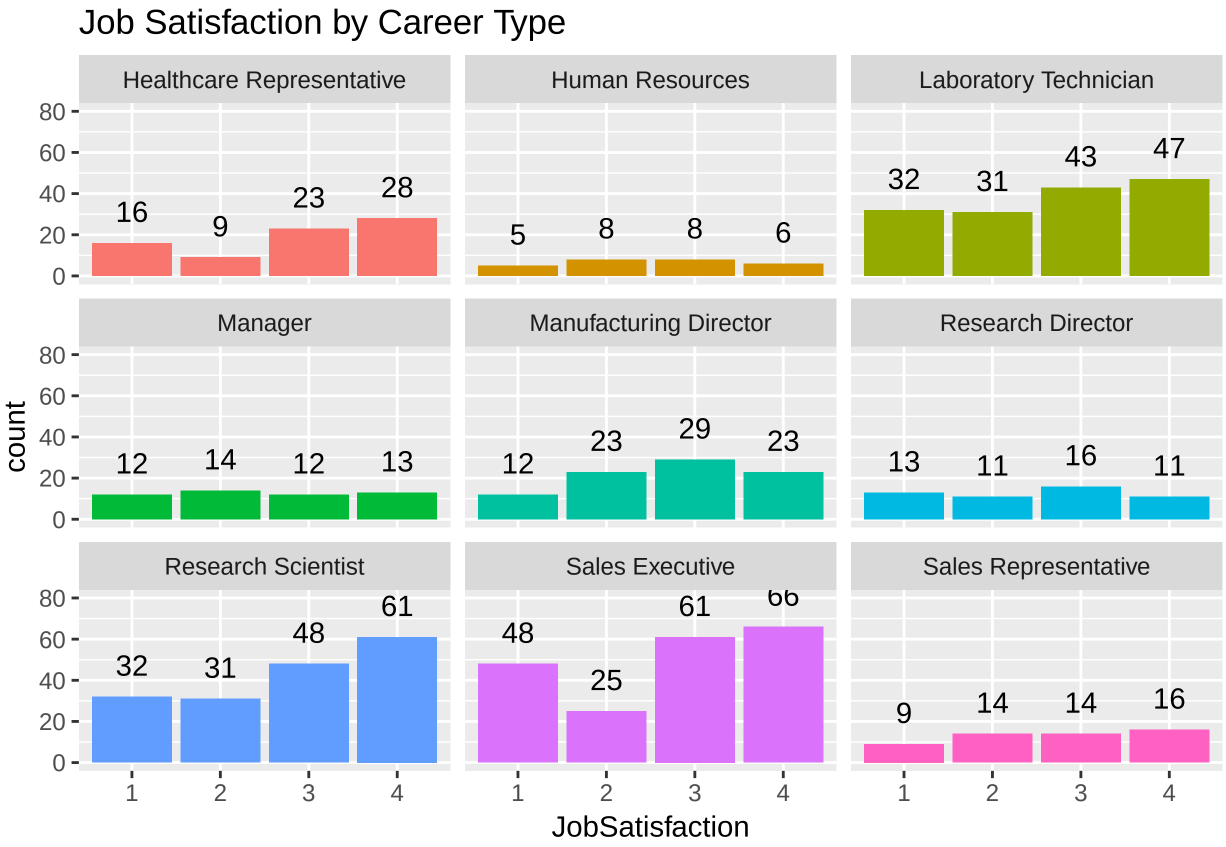
Quantiative Feature Distributions

Below are the distributions of the quantitative variables in the data. All numeric variables in this dataset are discrete however, several variables such as `NumCompaniesWorked` `PercentSalaryHike` `YearsSinceLastPromotion` have a small number of "levels". These features will be converted to factors for model use. The other quantitative variables have many values and will be left as numeric. The engineered feature `raise`, `totalYears`,`MonthlyIncome` are right skewed. These observations will be left as is as the models in use are flexible enough given the central limit theorem and the large sample size for these features. Several of the features such as `HourlyRate`, `EmployeeNumber` are more uniform in nature but more normal compared to the skewed distributions.



Job Satisfaction

Job satisfaction is an important feature in the `Attrition` classficiation models. Higher job satisfaction will be shown to have a positive impact on lower attrition. `JobSatisfaction` is defined as 1 - "Low"; 2 - "Medium"; 3 - "High"; and 4 - "Very High." When looking at `JobSatisfaction` by career group, Research Scientist and Sales Executives have the most satisfied individuals with 35% and 33% of their respective populations reporting "Very Satisfied." Research Directors have the largest perentage of dissatisfaction at 25% of their respective population reporting "Low" job satisfaction.



Classification with Logisitic Regression

Logistic Regression: Model Selection

Traditional forward, backward, stepwise model selection will be used to identify the best model for cross validation. This will be done by comparing selection method AIC or residual deviance where the lowest values are the most desirable. Using this method prior to cross validation helps generate models with better prediction ability. Below the model performance statistics for each selection type are shown. Based on this table, the stepwise selected model is the best candidate for cross validation with the although it's AIC is larger than the forward selected model, its residual deviance is almost 4 units less than the forward model's residual deviance. This suggests that the forward model has a tighter fit to the data.

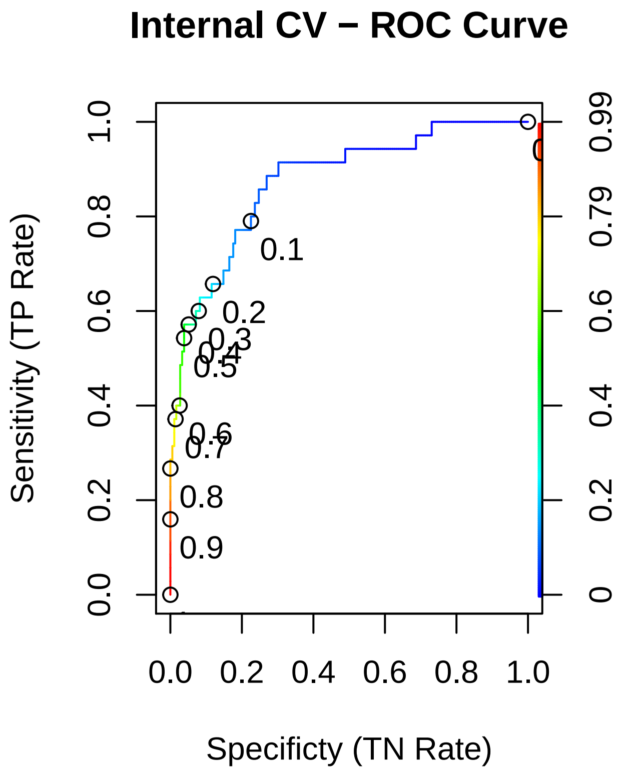
Selection Type	AIC	Residual Deviance
Forwards	518.46	430.46
Backwards	518.95	426.95
Stepwise	518.95	426.95

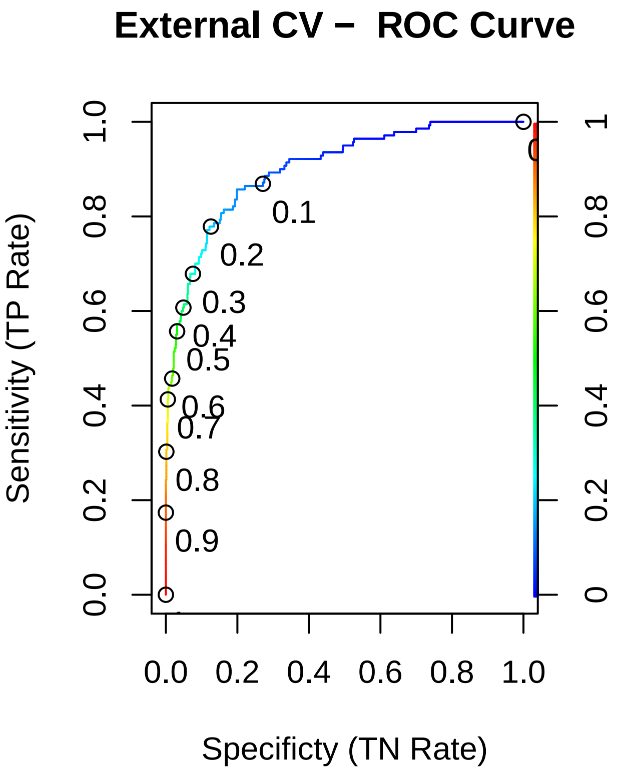
Logistic Regression: Internal Cross Validation and Model Development

Internal cross validation is used to tune a model for better predictive ability. The data will be split into a 75% training set and a 25% validation set. The training set will be divided into 25ths where the model will be fit to each partition and the process will be repeated 5 times. At the end of this process the final model will be generated and will be evaluated using confusion matrices and external cross validation.

Threshold Tunning

ROC/AUC Curves can be used to assess model performance and to assist in tuning the cutoff threshold for the response variable. Doing so can improve sensitivity or specificity, but not both. Setting and adjusting the threshold at which an observation is classified into one group or another may be beneficial as long as an ROC/AUC curve is referenced an the adjustment it not too extreme. In the case of the internally validated forward selected model we can see in the "Logistic Regression: Internal and External Cross Validation Model Performance Comparison" section that adjusting the threshold from 0.5 to 0.45 improves specificity while marginally affecting positive/negative predictive values and sensitivity.





Logistic Regression: Internal and External Cross Validation Model Performance Comparison

Several statistics are used to measure model performance. **Accuracy** is the total number of correct predictions; **precision** is the proportion of correctly identified positive observations and all predicted positives; **Sensitivity** is the ratio of predicted positives over true positives; **specificity** is the ratio of predicted negatives and actual negatives; and the **misclassification rate** are the total number of misclassified observations over the total number of observations. The AUC value for both models are high, this suggests that the models will have good predictive ability. When comparing the internal and external CV statistics we can see that both models perform similarly. Both have sensitivity and specificity rates that approach 60% and both have high accuracy and precision.

FittedModel:$Attrition = Age + BusinessTravel + DailyRate + DistanceFromHome + EmployeeNumber + EnvironmentSatisfaction + JobInvolvement + JobLevel + JobRole + JobSatisfaction + MaritalStatus + NumCompaniesWorked + OverTime + RelationshipSatisfaction + StockOptionLevel + TrainingTimesLastYear + YearsSinceLastPromotion + totalYears + overtimeRate$

	Internal CV Statistics	External CV Statistics
Accuracy	0.88940	0.89655
Sensitivity	0.95055	0.59346
Specificity	0.57143	0.83908
Precision	0.92021	0.89655
Misclassification Rate	0.11060	0.11060
Area Under the Curve	0.87535	0.90186

Naive Bayes Classification

Naive Bayes: Model Selection

Naive Bayes is an algorithm that utilizes Bayes theorem to classify observations. For this model variables were selected by hand and tested. Features derived from the initial exploratory data analysis were used for this model. During the initial training of the model cutoffs were set at 0.4 so that specificity and sensitivity metrics could be met. Laplace smoothing is used to smooth categorical data, using this technique aids in model performance.

Naive Bayes: Model Performance

As with the "Logistic Regression: Internal and External Cross Validation Model Performance Comparison" section model performance statistics such as: accuracy, sensitivity, specificity, and precision. These statistics are used to evaluate the model performance where the target specificity and sensitivity is at least 60%. For the Internal and External CV have sensitivity and specificty above 60%,

	Internal CV Statistics	External CV Statistics
Accuracy	0.84943	0.88904
Sensitivity	0.88904	0.64286
Specificity	0.64286	0.92847
Precision	0.92847	0.84943

Classification Model Selection

Based on the fit statistics generated by the Logistic Regression Model and the Naive Bayes Model, the Naive Bayes Model out performs Logisitic regression in Accuracy, Sensitivity, Specificity, and Precision. Please refer to the "Naive Bayes: Model Performance" and "Logistic Regression: Internal and External Cross Validation Model Performance Comparison" sections for the fit statistics. Based on the evidence the Naive Bayes model will be used for prediction.

Linear Regression

Linear regression will be used to predict the Monthly Income of job candidates. Features are derived from the exploratory data analysis section where features were created, removed, or modified. Please see the "Exploratory Data Analysis" for the full process.

Linear Regression: Model Selection and Training

Linear regression models were selected in a similar manner as with logistic regression. Please see the "Logistic Regression: Internal Cross Validation and Model Development" section on the full process. As with logistic regression forwards, backwards, and stepwise model selection was preformed to select the best performing model based on RMSE, adjusted R-squared and their p-values. Once selected that particular model will be cross validated. The main difference between these two models is that Linear Regression predicts continuous quantitative variables, whereas logistic regression is designed to classify dichotomous variables.

Linear Regression: Model Performance and Validation

The stepwise selected model was trained in cross validation using Ridge regression where the test set was partitioned into 25 sets and ridge regression was performed on each partition. This is repeated 10 times before the final model is generated. Ridge regression is a technique that helps correct for multicollinearity by adjusting the standard errors of each feature. To do this the technique introduces bias for correction. When comparing the RMSE of the initial model to the cross validated model both models have RMSE's that are approximately 720 with an r-squared of at least 97%. These consistent fit statistics suggest that the model will perform well in prediction.

Statistic	Initial Stepwise Model	Cross Validated Model
RMSE	723.9	742.3
Multiple R-squared	0.9758	0.9742

Linear Regression: Fitted Model and Interpretation

Based on the model fit statistics, both training and cross validated models perfromed well. Below is the model generated by training and validation. The top three contributors to `MonthlyIncome` are: `JobRole`, `JobLevel`, and `PercentSalaryHike`. For `JobRole` compared to a Health Care Representative (while holding other slopes constant), a Research Director sees an increase of $1861.96 a month, with a standard deviation of $164.42; this level of `JobRole` is significant with a t-value of 11.341 and p = 0. The second largest contributor to `MonthlyIncome` is `JobLevel` - when moving from Level 1 to Level 5 (while holding other slopes constant) there in an increase of $6335.30 in monthly pay with a standard deviation of $286.31; this level of `JobLevel` is significant at a t-value = 22.128 and p-value = 0. The thrid largest contributor to `MonthlyIncome` is 'BusinessTravel. Compared to someone that does not travel (while holding other slopes constant), a business traveler will make $198.75 more a month with a standard deviation of $79.99; this is also signficant with a t-value = 2.484 and p = 0.013165. Since all these feature levels are significant it suggests that their values are greater than or less than zero. This means that these feature levels may infact contribute to a higher monthly salary. When looking at the intercept, the average individual will make $4542.78 a month when all other slopes are 0.

Fitted Model:$MonthlyIncome = BusinessTravel + DailyRate + Gender + JobLevel + JobRole + NumCompaniesWorked + PercentSalaryHike + YearsSinceLastPromotion + totalYears + raise$

Conclusion

The models generated provide good predictive ability for attrition and salary trends. Exploratory data analysis allowed for feature engineering to help improve model performance. For the classification models, thresholds could be adjusted to determine at what probability an observation would be considered "Attrition" or "No-Attrition." The Naive Bayes model out preformed the logistic regression by a small amount in the specificity and sensitivity range, but both models were comparable in accuracy and precision. For the linear regression model, we were able to identify that `JobRole`, `JobLevel`, and `BusinessTravel` contributed the most to an individual's monthly pay.

Presentation Video

Please see the attached presentation for DDSAnalytics findings HERE: https://youtu.be/Afq9FAA5lKc