Final Project

The purpose of this final project is for you to apply what you have learned during the class.

The data set is simulated HR data and I want you to build a model that will predict who is going to leave.

You are being given the existing data and I will release the new data for your model to be tested on as we get closer to the deadline.

Below are the objectives of this project:

* Produce tables and graphs (ggplot2) for each of the variables
* Produce tables and graphs for some crosstabs of the variables (at least 3)
* Create at least 4 models that will be evaluated on the following:
  + Confusion Matrix Performance
  + AUC performance
  + Simplicity
  + Interpretability
* Write up the results as you would for yourself or a fellow data analyst who either needed to reproduce it or build off of it under a tight deadline.
* Write up the results as you would for a manager or other lay person on actionable items to decrease turnover and the expected monetary savings
  + Assume the cost per person leaving is $100,000
  + Assume the intervention is $5,000
  + Assume the intervention works 50% of the time
  + Remember that if you determine the model has identified someone who will leave and you administer the intervention and it works, you will have saved $95,000 ($100,000 saved minus $5,000 for the intervention)

Using the provided data which included variables such as employee satisfaction levels, last evaluation score, number of projects, average hours worked in a month, tenure, promotions, and salary, we sought to find a model that could accurately predict which employees would leave the company. Given the variables in the data, our goal was to determine which predictors may provide the most signal that an employee intended to leave. If this information was known, we may be able to intervene and prevent the employee from leaving, thereby saving the company money.

For this objective, it has been determined that the loss of an employee costs the company $100,000. We are recommending an intervention program with a cost of $5,000 per employee. Based on our research and data, we expect the intervention program to work approximately 50% of the time. With these costs in mind, it is understood that the company will save $95,000 per employee who does not leave.

FOR DATA ANALYSTS

The following code was used to run and interpret 4 models to predict attrition using the dataset. The first sections include exploratory data analysis and visualizations to understand the context of the data. The data was relatively “clean” with no missing values or duplicated observations. Models were developed using `tidy` modeling principles, including workflows and recipes. The recipes include preprocessing information to normalize and balance the data. Results of each model included:

Support Vector Classifier - accuracy

Logistic Regression - accuracy

Random Forest - accuracy

Neural Net - accuracy