**Description**:

DDSAnalytics has requested an analysis of employee data with the intent to decrease the rate of attrition, as well as form a model for predicting attrition, and another model for predicting monthly income based on whatever factors are found to be influential. Additional analysis is welcome.

Analyze existing employee data at DDSAnalytics to:

1. Predict employee Attrition
2. Identify any Job-Role specific trends
3. Determine the top three most influential factors that contribute to Attrition
4. Establish criteria for predicting monthly income, and create a regression model

**Process:**

First, the data was cleaned and organized. There were no missing values, and the data was comprised of both categorical and continuous variables. The categories “EmployeeCount”, “Over18”, and “StandardHours” all only had one level so they were removed from the data.

Next, the mean of each continuous variable, and a frequency bar chart for categories, were plotted according to Job Role. Sales Representatives are, on average, the youngest, live closest to work, the lowest monthly income, fewest working years, and fewest years at the company. When looking at attrition, you can see that they also have the highest rate of departures compared to any other job role (almost 50%). If you take a look at business travel, a substantial percentage of sales reps travel frequently when compared to other roles. Sales representatives make up a relatively small portion of the Sales department, especially when compared to Sales Executives. This may be due to the higher rate of attrition in that role, but at a lower income it seems like you might get value at accelerating hiring of this role.

Environment Satisfaction scores were mostly evenly distributed within and across every job role. You also look to have good gender diversity across all job roles, although research scientists and lab techs look to be slightly higher percentage male, as does human resources.

The data was heavily skewed toward “No” for attrition, so it was randomly subsampled to compare with less bias. A best seed was found for creating test and train samples, and Naïve Bayes was run on the data. It was found that the top three highest probability Attrition factors are employees who Travel Rarely (conditional prob = .68), Stock Option level of 0 (cp = .71), and performance rating of 3 (cp = .85). The model we created had an accuracy of .833, Sensitivity of .871, and Specificity of .800 when run against the test set. When used to predict attrition for the competition data, the outcome was again heavily skewed toward “no” attrition. So if that data sample is representative of our starting dataset, this reinforces our model.

For our second model, we set out to determine what factors can be used to predict monthly income. Again, the data was split into a training and test set, and this time we ran linear regression against all variables. An ANOVA was run to determine what variables are significant predictors in our model. We then re-ran the regression, and ANOVA until all our predictors had significant p-values < .05. Age, Attrition, Business Travel, Department, Education, EducationField, EnvironmentSatisfaction, Gender, JobInvolvement, JobLevel, and JobRole were all deemed significant in our model. Each had a p-value <.001 except for business travel with a .008. The model was run against the test set, the train set, and the full dataset to receive respective RMSE values of $1200.83, $973.09, and $1064.57. We then created residual plots for each of these, and see that the data is randomly scattered along the zero line with no influential points. Finally, density plots were compared between the predictions on the competition sets, and on the downsampled data. They have very similar shapes, which reinforces the strength of the model.



















