Predicting IBM Employee Attrition

Aiman Chughtai

Abstract

The goal of this project was to propose a classification model to predict whether or not an employee would churn, use those predictions to target employees based on how "at-risk" they were for attriting, and suggest recommendations on how to reduce the chances of losing the employee. I used a dataset found on Kaggle.com of fictional "IBM HR Employee Analytics" (https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset) . After refining the model, I built an interactive dashboard to visualize and communicate my results using Tableau.

Data

The dataset contains 1470 employees with 35 features for each, 9 of which are categorical. The target variable was Attrition (Yes/No), which marked whether or not an employee had churned. Some of my important features included age, job level, monthly income, department, job role, level of stock options, total working years, years at current role, years with current manager, and job involvement. While many features were highly correlated with one another, due to the nature of employee analytics, I believe that is to be expected- no one feature is indicative of why an employee might leave an organization, just as no one feature describes how an employee is compensated salary-wise.

Tools and Algorithms

- Microsoft Excel
- Python
- Numpy
- Scikit-learn
- Seaborn
- Matplotlib
- Tableau

Logistic regression modeling and evaluation

Logistic regression fitting was performed by SciKit-Learn. Model was evaluated with a confusion matrix and performance metrics (accuracy, f1 score, precision, recall). An ROC curve was plotted to view AUC.

- Accuracy 0.897
- F1 0.596
- precision 0.8
- recall 0.475

Results/Design:

I began with the IBM HR Employee dataset i obtained from Kaggle. I imported it into python in order to quickly clean the data (fill nulls if any, remove duplicates). I then exported that dataframe as an excel file loaded it into excel in order to do the remainder of my EDA. I was interested in examining each feature's relationship with the target (attrition), which I did by creating Pivot Tables. I created preliminary visuals in excel in order to guide my advanced visuals in Tableau. Once I had a general idea of important features from my eda in excel, I went back to python to view correlations using seaborn heatmap and to create my logistic model using scikit learn. I tuned hyperparameters using GridSearch, got my accuracy score as well as other metrics to gauge accuracy of my model such as f1-score, precision, recall and visualizing a confusion matrix. Finally I made an ROC curve to view AUC. Once I understood feature contribution to target, I created graphs in tableau in order to visualize those findings.

Communication

A presentation, slides, and tableau dashboard