Case Study: Hewlett Packard & Employee Attrition

**Overview**

This case study describes a predictive analysis project by Hewlett-Packard in which the company attempted to predict employee attrition by calculating employees’ “Flight Risk” scores. Hewlett-Packard, better known as HP, is an information technology company. It is renowned for its hardware components such as computers and printers, but it also provides software and other technology services for customers. HP is quite a profitable company, having a total revenue of $127 billion in 2011. However, with such impressive financial gains, it requires talented employees to build and power the products. Their 330,000 employees provide a source of data for understanding and deciding employees’ risk of leaving the company and increasing attrition rate.

For the sake of having a dataset for this case study, employee attrition data can be found at this link ([Dataset Link](https://github.com/phillipseb/DSC530/blob/e8fed218e0f80af1f6ad8b709e554e26a1e3ac13/ThinkStats2-master/code/WA_Fn-UseC_-HR-Employee-Attrition.csv)), and it will be referenced throughout the document.

**Business Understanding: Defining the Problem**

For the predictive analysis project, training data was built around the 330,000 employees at Hewlett-Packard. This data included two years of information on employee salaries, raises, job ratings, and job rotations. Along with these measures, there was demographic data collected on the employees as well, such as age and gender.

The dataset also contained a simple binary categorical variable which denoted whether employees in the training set had quit.

***Defining the Target Variable***

As mentioned in the overview section, HP’s goal was to calculate “Flight Risk” scores for their employees to be able to aggregate and rank the employees at greatest risk for quitting their careers at the company. Therefore, the target variable for modeling are these scores.

**Data Understanding**

The employee attributes are shown in Figure 1, along with their data types before any transformations or casting.

Figure

Table

Description automatically generated

There were no duplicates or missing values found in the data!

There are also variables in the dataset which pertain to employee ratings of some concept such as job satisfaction, work life balance, etc. The employee rating values range from 1-5, like what would be provided on an employee survey. Majority of these rating values were in the 3-5 range, which could indicate that employees are relatively satisfied with their careers at HP, or it could be indicative of skew with survey results given that sometimes they are not taken seriously by users. However, from understanding of the data, the responses to the survey topics seem to be accurate reflections of the employees’ feelings. No corrective action was taken for this.

**Data Preparation**

There were some variable transformations needed though to derive more detail from the original data. As seen in the dataset, there is a variable for ‘MonthlyIncome’. However, usually with employee information, it is more useful to have the income as a yearly salary. Therefore, the ‘YearlyIncome’ variable was derived from ‘MonthlyIncome’ by multiplying its values by 12 to get the proper amounts.

In terms of outliers, since employees experience quite a range of salary opportunities and job changes, there were no significant outliers that needed to be removed as they all tell a story in employee experiences and how it could possibly affect attrition. The goal with the data is to learn which combinations of factors define the type(s) of employees most likely to quit their jobs.

One-hot encoding was performed on the ‘Attrition’ variable to convert the categorical “Yes” and “No” for employee quitting into 0’s and 1’s. Also, to get all numerical values on the same scale, feature scaling methods were applied as well to normalize the data values and apply appropriate weights for numerical representation. Variance-Inflation Factor (VIF) was also used as the feature reduction method, to only select the features which were not highly correlated with each other for final use in the model.

**Modeling**

Out of the training set of data, the predictive model was tested on a specific team at HP, the internal Global Business Services (GBS) team. The group had about three hundred employees, and it had been showing a high attrition rate of up to 20 percent, making the team a great candidate for flight risk predictions.

The actual model applied on the two sets of data was a decision tree analysis, which allowed for the output of sets of decision trees with randomly selected predictor variables. Therefore, by having a larger number of variables in the dataset, it allowed for the decision tree analysis to be relatively accurate given the large number of variables to select randomly. The results of the models were analyzed via confusion matrices to determine independent accuracy. The best fit model was then applied on the test data from the GBS team, and then probability or flight-risk scores were calculated for each employee in the group.

***Model Interpretation***

From the model outputs and calculated flight-risk scores, the importance of certain variables in the dataset became clearer. Some of these include employee salary, number of raises, and performance ratings. Their influence on decreasing Flight Risk came from employees having higher salaries, more raises and increased performance ratings. Rotational opportunities also played an important role in decreasing risk of losing employees, as they had more chances at the company to try out roles and gain new skills.

There were also variables that had a reverse impact on the Flight Risk in causing an increase rather than a decrease. For example, with getting a promotion, if the promotion did not also lead to an increase in salary, this caused an adverse impact on flight risk rather than benefitting it.

**Deployment**

Once the predictive Flight Risk model was applied company-wide at HP, it identified “$300 million in estimated potential savings with respect to staff replacement and productivity loss across all HP employees throughout all global regions” (Siegel, 2016). Also, for accuracy in predicting employees that would quit, “the 40 percent of HP employees with higher Flight Risk scores included 75 percent of the quitters (a predictive life of 1.9)” (Siegel, 2016).

**Summary and Conclusions**

For this business problem of predicting employees’ Flight Risk scores, decision tree analysis was utilized to identify the variables which had the greatest impact on increasing or reducing employee flight risk. Scores were also calculated for each employee to aid in identifying top employees which needed more attention and intervention from designated management. Hewlett Packard deployed this model on their employees with great caution to ensure that its results were in the right hands and were not being maliciously used. Employee characteristics are masked in the score reporting system to ensure confidentiality as well.

HP has seen a decrease in attrition rates from 20% to 15% for the GBS test team since the business problem was attempted to be resolved. The attrition rates continue to decline, and hopefully will continue to as the model is improved and re-deployed, as well to a larger population.

**References**

1. Siegel, E. (2016). Chapter 2: With Power Comes Responsibility (Ethics). In *Predictive analytics: The power to predict who will click, buy, lie, or die*. essay, Wiley.