Case Study: Predicting Flight Risk at HP

Bilal Kudaimi

Overview

A company is much like a sailing ship. Everyone has a role to fulfill and must do their part to keep the ship afloat. However, a company is not a fully sealed ship; there is always the risk that some of its most talented employees may jump ship and leave, a concept known as flight. It is estimated that employee flight incurs a cost to a company between 1.5-2 times the lost employee's salary for hiring and training, and this isn't including the costs of lost productivity due to a vacancy and lost talent. Thus, many companies have been focusing on employee retention, including the tech giant Hewlett-Packard (HP).

This case study describes a predictive analytical project to reduce employee flight among employees of HP, a multinational technology company that manufactures and sells various electronic devices and IT solutions. Despite being a tech giant in its field, it still suffers heavily from employee flight, with some of its departments reporting as much as a 20% turnover of employees. Due to this high percentage and the costs associated with many employees leaving, two HP employees, Gitali Halder and Anindya Dey, decided to undertake a project to build a machine learning model that could accurately assign a "flight risk" to each employee, essentially predicting whether an employee would leave or stay with HP. By predicting which employees are most likely to leave, HP can focus its resources on retaining those employees and reduce some of the high costs associated with employee flight; this project was estimated to save about $300 million total.

Business Understanding

The main business problem is that many departments within HP have such a high number of employees leaving HP, whether it be to other companies or just leaving outright. Halder and Dey remedied this problem by constructing a predictive analytic model that could accurately predict which employees would leave within a specific time period. To do this, they compiled two years' worth of employee information on more than 330,000 HP employees into one dataset. They then tacked on whether each employee left the company to generate a dataset usable by a machine learning model.

The target variable, or the "y" variable used to train the predictive model, was in the binary column called "Left the company." Elements of this column were either a "YES" if an employee left or "NO" if they still work for HP. The objective was to build a model that draws a relationship between the dataset's target variable and the rest of the dataset's features to predict two things: which employees were most at risk of leaving HP (via the flight risk score), and which features would be most likely to retain an employee who is at risk of leaving HP. The criteria for successfully building such a model would be to generate high model evaluation metrics such as accuracy, precision, recall, etc. If such criteria were met, the model could be deployed.

Data Understanding/Data Preparation

The dataset used to build the model was compiled by Halder and Dey using information they obtained on all 330,000 HP employees at the time. This information included the employee's salaries, number of raises, raise amount, job performance

ratings in multiple categories, number of job rotations, and whether the employee left HP or not. After compilation, the dataset had very few missing values, as HP keeps track of all employee salaries, raises, and performance ratings. Thus, no missing data imputation was conducted.

Since all features of this dataset were numerical, it must be certain that each feature had matching number types. For example, the salary column must contain all floats (any integer values must be converted to floats) and the number of rotations column must contain all integers (any float values must be converted to integers). The numbers' data type will be checked and made to match within each column. Lastly, the binary column of whether an employee left must be converted to integers instead of just "YES/NO" so the model can parse the column and draw its predictions.

Modeling

The data must first be split into training and testing data to build a logistic regression model. Logistic regression is usually used when the target variable for prediction is binary, so it will be used to predict employees' flight risk. I will use 80% of the data for training and 20% for testing to ensure that my chosen scoring metric, accuracy and precision, remained low. Accuracy and precision are ratios constructed using the true/false positive and negative values from the model; positives here would be if an employee will leave. With the split data, I would fit the model to the training data using a 10-fold cross validation then report the resultant root mean squared error. I would then pass the fit model to the "x" testing data and compare the output target

variables to the actual values of the target variable to see if the output error is reasonable compared to the predictions.

There is, however, one question: If the target variable is binary, then the model should also output binary variables, so how then, could it output a numerical flight risk score? Logistic regression output a probability float that was converted into a binary variable. In this case, the model output the probability of an employee leaving HP – the flight risk score – which was then converted into a binary variable (whether the employee will leave) using a cutoff threshold (e.g., above 70% risk the employee is considered a YES for leaving). For this model's cutoff threshold, I would use a value around 60% to account for margins of error in prediction.

Deployment

Predicting HP employees' flight risk scores did not come without a slew of privacy concerns. What would happen if everyone saw these scores? Would everyone, or only the employees, resent HP and the Halder-Dey team? Would news of this model ironically cause many to resign their positions? This is why the logistic regression model for this case was deployed with extreme caution. Only a select few managers who were trained to understand the ramifications of these scores were allowed to view the flight risk scores, and even then, only for the employees under them, as well as explanations of the scoring reason. The scores were kept decrypted and only said managers were given decryption keys.

Conclusion

The logistic model resulted in a low revealed many expected trends about flight risk, but also a few unexpected ones. For example, while it was shown that employees with higher salaries, raises, and ratings quit less, a greater amount of job rotations also reduced flight risk. This is most likely because boredom is a cause in employee flight and a job is kept more interesting with more changes. Furthermore, the number of promotions decreased flight risk in all teams except the Sales Compensation team, where only the addition of a significant pay hike served to decrease the risk.

Based on these findings, the model has the potential to benefit other companies in the future and not just HP. However, there are several implications to consider when implementing flight risk prediction models. Is the data used for it obtainable in a legal manner? Is such a model ethical? What are the consequences of false positives and negatives (i.e., misidentifying employees as loyal/disloyal)? While the jury is out on whether flight risk scores are inherently unethical, if managed correctly, flight risk models will greatly benefit a company without causing employees to feel they are being accused of disloyalty.

References

Abbott, D. (2014). *Applied Predictive Analytics – Principles and Techniques for the Professional Data Analyst.* John Wiley & Sons.

Heinz, K. (2020). *Dangers of Turnover: Battling Hidden Costs.* Built In. Retrieved on July 1, 2021 from https://builtin.com/recruiting/cost-of-turnover.

Siegel, E. (2016). *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die.* John Wiley & Sons.

Srinivasan, S. (2019). *Business and Data Understanding in Data Science Lifecycle.* Medium. Retrieved on July 1, 2021 from https://medium.com/@srivatsan88/business-and-data-understanding-in-data-science-lifecycle-58f8e0588c66.