Machine Learning Engineer Nanodegree

Capstone Proposal

Kim Alderman
December 9, 2016

Employee Turnover Prediction

Domain Background

Machine learning (ML) is being applied to many facets of business, from supply chain management to customized/targeted marketing plans. One business area that is gaining traction in the use of data science more broadly is human resource analytics (HR analytics). Data mining was successfully used to evaluate applicants (Chien & Chen) and Harris showed a creative use for machine learning in HR by using incentivized crowdsourcing to review job applicant resumes. There has also been research into using decision trees to assist HR professionals in classifying talent among existing employees and identifying potential targets for promotions.

Even with the current research, there is significant work to be done to fully realize the benefit of HR analytics, both because of the of the potential benefits and the complexity of the problem. It is costly to attract and hire new employees, and employee performance can help drive some key performance indicators (KPI), so HR professionals would benefit from models to improve hiring and retention. However, unlike much of the numerical data streams utilized in other departments like finance and purchasing, HR often relies on more subjective data sources. It is probable that machine learning could be used to uncover hidden factors that improve these prediction models.

I personally became interested in this machine learning application through a rather serendipitous circumstance. A colleague in the quick-serve restaurant (QSR) industry recently hired me to perform an analysis of an employee engagement survey his company had completed. In researching relevant studies for these types of surveys, I found that data analysis (and machine learning) were becoming recognized as important tools for HR. Further, I was fortunate to be granted a scholarship to attend the inaugural O'Reilly Artificial Intelligence Conference (AICon) in New York in September 2016. While there were many discussions around using ML and AI for applications like autonomous vehicles, object recognition, and chat bots, there was virtually no attention to HR issues. Thus, I feel this is an area ripe for further research.

Attracting and training a new employee is costly for companies. Initially, the new employees contribute little to overall productivity or profitability. Over time though, the hope is that there will be a breakeven point, and eventually those upfront costs will be worth it as the employee starts to create value for the company. Therefore, it is important to reduce attrition, particularly of high-performing employees. In researching possible datasets for this project, I found an open dataset on

Kaggle that includes numerous employee features that could be useful for creating a retention model.

Problem Statement

The problem is identifying the high-performing employees that are most likely to leave so that HR can intervene to possibly retain them, thereby reducing turnover of these high-value employees. The machine learning algorithm will be used to predict classification probability for leaving the company. The inputs will be employee data features such as satisfaction and time spent at the company (see next section for detailed discussion of inputs). This classification model can them be used on existing employees to identify the highest risk, high performers, allowing HR to focus retention efforts.

Datasets and Inputs

The data was obtained from a Kaggle dataset (https://www.kaggle.com/ludobenistant/hr-analytics), which was release under a CC BY-SA 4.0 License. Input features of the dataset include: employee satisfaction, last evaluation, number of projects, average monthly hours, time spent at the company, work accident (yes/no), promotion in last 5 years, sales, and salary. The dependent variable is the data field indicating whether the employee has left the company.

Employee satisfaction is expressed on a scale of 0-1, with 1 representing highest satisfaction. It seems logical that this may be correlated with retention – unsatisfied employees may be more likely to leave the company. Last evaluation is the overall "score" given to the employee at their last annual evaluation on a scale of 0-1, with 1 representing the highest performers.

Number of projects, average monthly hours, and time spent at company are all expressed in integer form, although average monthly hours will need to be normalized since they are on a different scale than the rest of the features. It will be interesting to see if these features correlate positively or negatively with attrition. Work accident and promotion within last 5 years are binary features (1 yes/0 no). Sales and salary are categorical features denoting the department (i.e. Sales, Accounting, etc.) and salary tier (low, medium, or high).

Solution Statement

Companies routinely collect data on employees, from the hiring phase all the way through when an employee leaves the company. This project will attempt to build a model to predict employee retention given some of that data as input features. This will be a supervised learning project as the observations (i.e. employees) are labeled with their employment status. The metric will be the accuracy of the model in predicting if the employee is still with the company or has left.

In addition to providing a predictive model of which employees are likely to leave, I hope to gain some insight to provide prescriptive analytics. By observing correlations between high-performing employees (as measured by the last evaluation feature) and the likelihood of leaving, we may be able

to identify high-value targets for the company to work with for retention. If a company could know in advance the signs that a high-performer is at risk for flight, concrete actions might be used to increase chances for the employee to remain. Also, because the features used for this project are collected routinely by human resources, it should be possible to use new and updated data to increase the accuracy of the model.

Benchmark Model

Based on initial exploratory data analysis, 23.81% of the observations (employees) have left the company's employment, with the remaining 76.19% remaining with the company. The default prediction would therefore be to essentially guess that the employee had not left, yielding an accuracy of classification of 76.19%. My model needs to surpass this accuracy to be of any practical use.

Evaluation Metrics

While I had originally proposed that the evaluation metric should be classification accuracy, a Udacity reviewed pointed out that because the data is imbalanced (more observations of employees staying than of leaving), simple classification accuracy would not be appropriate. Instead, the metric will be F1-score which is a weighted average of precision and recall (https://en.wikipedia.org/wiki/F1_score). Precision is the number of correct positive predictions (true positive, TP) divided by the number of all positives (TP+FP, where FP is false positive). Recall is the number of correct positive predictions (TP) divided by the number of actual positive observations (TP + FN, where FN is false negative). The F1-score is then calculated by:

$$F_1 = 2 \cdot \frac{1}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Project Design

The project will be broken into four key blocks: data exploration, data cleaning, model building and testing, and result presentation.

Data Exploration:

Fortunately, the data has already been obtained and can be downloaded at https://www.kaggle.com/ludobenistant/hr-analytics During this initial phase, I will be getting familiar with the dataset by performing summary statistics, data visualizations, and looking for correlations between features.

Data Cleaning:

In this phase, I will determine how to best handle missing data, outliers, encoding of categorical variables, and normalization of features.

Model Building

Several algorithms may be employed to find a model that improves upon the baseline model. Initially, logistic regression and decision trees will be used to discover which features, if any, are most predictive of an employee leaving. The coefficients of the logistic regression and the top level nodes of the decision tree will be used to identify those features which are most relevant. An advantage of these models is that they may yield actionable insight that the company can use to reduce attrition.

To further improve the accuracy, additional algorithms may be used. Most likely candidates would be random forest and support vector machines.

Each implementation will be run with a Python script and kept in separate files to ensure replicability.

Works Cited

Chien, Chen-Fu, and Li-Fei Chen. "Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry." *Expert Systems with applications* 34.1 (2008): 280-290.

Harris, Christopher. "You're hired! an examination of crowdsourcing incentive models in human resource tasks." *Proceedings of the Workshop on Crowdsourcing for Search and Data Mining (CSDM) at the Fourth ACM International Conference on Web Search and Data Mining (WSDM)*. 2011.

Jantan, Hamidah, Abdul Razak Hamdan, and Zulaiha Ali Othman. "Human talent prediction in HRM using C4. 5 classification algorithm." *International Journal on Computer Science and Engineering* 2.8 (2010): 2526-2534.