ML Nandodegree Proposal

December 9, 2016

1 Proposal: Predicting Why Employees Leave

1.1 Domain Background

According to the Bureau of Labor Statistics the median number of years that wage and salary workers had been with their current employer was 4.2 years in January 2016. While this number varies from industry to industry the story of an employee who sticks with one company for the entirety of a working life seems to be rather antiquated. This combined with the fact, that "employee turnover has been identified as a key issue for organizations because of its adverse impact on work place productivity and long term growth strategies". [1] One of the key issues with a high employee turnover rate, combined with but beyond cultural and sociological effects, is the cost associated with it. Research shows that the replacement cost for an hourly worker can be as high as 50 % of her annual salary. This number increases with the seniority of the skillset up to 200% for senior-level workers ans surges up to 400% for executive level positions.[2] It becomes obvious that the trend of shorter tenure in addition with the fact that a high employee turnover can be a costly endeavour.

Therefore it gets increasingly important to acquire the necessary tools for employers to understand where its workforce is standing. Additional insights from employer reports, scorecards as well as general statistical information can offer prediction values for companies when it comes to the longetivity of jobs. In this notebook we are trying to predict the likelihood of an employee quitting based on available information and are trying to offer clustering methods in order to gain actionable insights on how to prevent a high employee turnover.

1.2 Problem Statement

The problem to be solved is detecting the key elements of employee tenure and predicting whether an employee might be quitting her job. This problem affects companies from all sizes. Although it can be said that based on statistics there are certain industry traits that have a higher tenure time, relative speaking the tendency to switch jobs has increased over time. The boundaries of this problem of course are the underlying mechanisms of every individual company. It can not be generalized as in how to prevent but more as in how to predict. Company culture, direct reports, individual needs might turn out to be far more complicated than a list of features. However, if we fix this problem the path to a more successful and sustainable company culture is given. A desirable solution can be quantified by correct detection of potential job quitters and intervention opportunities created. It shall be stated that a mechanism such as this can only work as a supplement to human interaction and empathy skills.

1.3 Datasets and Inputs

Measuring employee satisfaction is a tough and highly interdepent task. There are a lot of different dimensions at play and turning them into quantifiably format (less to say machine-readable information) can pose a challenge. Information from employee reviews, tenure, balanced scorecards and key performance indicators can offer a first gateway to understanding an employee's desire to leave the company.

In a dataset published on kaggle we are offered information on current and former employees in concert with key features of their exmployment status. The dataset consists out of almost 15.000 data points and 10 variables. Unfortunately there is no available codebook besides a brief information on the available inputs. The inputs we will be using are:

- Employee satisfaction level
- Last evaluation
- Number of projects
- Average monthly hours
- Time spent at the company
- Whether they have had a work accident
- Whether they have had a promotion in the last 5 years
- Sales [describing the department]
- Salary

In addition the dataset provides an indicator whether an employee has left which we'll be setting as our dependent variable.

1.4 Solution Statement

We are trying to solve this problem on two dimensions by combining unsupervised and supervised learning algorithms. First we'll do an in-depth exploratory analysis in an attempt to explain the underlying relationships in the data set. Afterwards we'll be using clustering algorithms in order to create segments within our data set and enrichen the descriptive value of the data. Afterwards we'll implement a model that will empower employers to not onlyl detect, which of their own employees is about to leave the company but also determine which of the features contributes the most to a happy company culture. The success of our final solution will be measured by its predicting accuracy on a held-out test set.

1.5 Benchmark Model

A lot of research and work has been performed to retain talent and decrease employee turnover rates. From corporate funded research and analytics services of companies such as [3] to Management study classics such as [2] or more recent research [3] to name a few.

There are valid methods for predicting employee turnover rate and results that can be taken as benchmark models. In this research we'll be using the machine learning approach of Punnose and Ajit as a benchmark. Their research performed predictive tasks on company information with an AUC (Area under the Curve) of .86 on hold-out data as a best result using an XGBoost model.

1.6 Evaluation Metrics

AUC (**A**rea **u**nder the **C**urve) is a common evaluation metric for binary classification problems. It's value is between 0 to 1 and describes the accuracy of a binary classification based on its true

positive values. As described here, the area under the curve is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one.

To put this more into perspective, kaggle describes a way of viewing AUC as a plot of the true positive rate vs the false positive rate where the threshold value for classifying an item as 0 or is increased from 0 to 1. It can be stated, that if the classifier is very good, the true positive rate will increase quickly and the area under the curve will be close to 1. If the classifier is no better than random guessing, the true positive rate will increase linearly with the false positive rate and the area under the curve will be around 0.5.

1.7 Project Design

After describing the problem we are trying to solve and the data set we are going to use let's move over to the project design. In general the approach should be pretty straight forward and will be split up in following chronological steps:

1.7.1 Exploratory Analysis

To get a better grasp on the information that is included in the data set, we'll start off with a thorough exploratory analysis. After looking at the key metrics of the data we'll start asking basic questions regarding the status quo of our employees. We'll also try to figure out which features tend to be more describing than others when it comes to our prediction tasks and we'll be looking at correlations to discover underlying relationships in our data.

1.7.2 Data Preprocessing

From a short preview on the data set we can already tell that we are dealing with numeric but also categorical data. In order to make this information machine-readable we'll be applying a one-hot-encoding step. For numeric attributes we might consider normalization or calibrating. In addition, depending on our exploratory analysis, we might consider imputation techniques for missing data and outliers processing.

1.7.3 Deciding on algorithms & techniques

Even though we are already set on the type of machine learning we want to apply this will be the part where we're conducting additional research to find the best fitting algorithms for our deemed solution. There is an array of possibilities on how to tackle this problem and our key metrices to decide on a set of algorithms will be data-size, computational efficiency and cross-validation scores. However as mentioned earlier we'll be performing a clustering algorithm first in order to segment our employee data. This could be valuable especially when it comes to early detection and deciding on strategies for intervention. We will be using a 60/40 split upfront, to create a testing and a training data set.

1.7.4 Defining Models

We'll apply an array of tasks to our model in order to figure out the best way of handling our prediction task. This might include selecting, fine-tuning, and combining the best algorithms using techniques such as model fitting, model blending, data reduction, feature selection, and

assessing the yield of each model, over the baseline. To avoid overfitting and enable generalization we will be using cross-validation.

1.7.5 Discussion

Finally we'll discuss our model and our results in comparison to mentioned references and benchmark-models. This is, where we want to emphasize the general idea behind our approach and open it up to additional data sets from small to mid-size companies.

2 References

- [1] **Punnose R, Pankaj A** (2016) "Prediction of Employee Turnover in Organizations using Machine Learning Algorithms" in *International Journal of Advanced Research in Artificial Intelligence, Vol 5.*
- [2] **Weisbeck, D** (2015) "Fact or Hype: Do Predictive Workforce Analytics Actually Work?" on *visier.com*.
- [3] **Cotton J, Tuttle J** (1986) "Employee Turnover: A Meta-Analysis and Review With Implications for Research" in *The Academy of Management Review 11(1)*.