DNB Data Scientist for Enterprise Nanodegree Program

Capstone Proposal

John Gunnar Blostrupsen
November 22nd, 2018

Proposal

Domain Background

Nowadays, in all the hype on neural networks, artificial intelligence, machine learning and data science, it is important not to forget that there still is need for us humans to play along with the new technologies. For example, in the domain in which I have worked for the last 16 years (banking), the turnover of employees, especially the youngest ones, is accelerating. If a company really wants to stay competitive, it is crucial to know how to prevent key employees from leaving the company. Therefore, it is important to have insights into what are the main causes for employees to choose to quit. Such insight will help companies to keep more of their valuable employees. The conventional way to keep talents is increasing salaries and/or offer bonuses, but today's working environment demand other approaches. There are many opinions on this – one example can be found here at forbes.com.

There are quite a few articles to be found on the subject, and I would like to make a reference to this one at www.inc.com, where it is said that about 75 % of the reasons for costly voluntary turnover come down to things that managers can influence.

I would like to get further insight on this by using techniques I have learnt the last few months, attending Udacity's Data Scientist for Enterprise Nanodegree Program. (Suitable enough, attending this course has in itself made me more eager to stay with my current employer ☺).

Problem Statement

In an increasingly competitive job market, with continuous competition for attracting valuable competence and resources, it is key to not only attract the right people, but also keeping the employees you got. It is therefore important to aquire insight into why some employees choose to leave.

The reasons for employees to decide to quit may be many and complicated to comprehend, but there should be possible to extract some insight on this from data that contains information related to a company's employees, and I found a dataset on kaggle.com that I want to dig into, in order to try to understand reasons for employees to leave a company – focusing on finding eventual main reasons and try to make predictions on whether an employee will quit or not.

I consider this to be a classic classification problem, as I want to predict a label (whether an employee leaves or not), and I will also use the data given for each employee to find the main reasons for turnover, by investigating eventual correlations in the data.

Datasets and Inputs

The data I want to use in this project is this dataset I found on kaggle.com, containing data related to approximately 15,000 employees, including about 3,500 employees that already have quit and left the company. I chose this dataset because it looks pretty clean and tangible enough for me to conduct analysis on, as I am pretty new to Data Science and the implying techniques. The dataset contains both categorical and numerical data. Here is a brief overview of the features:

- Satisfaction level
- Last evaluation
- Number of projects
- Average monthly hours
- Time spent at the company
- Work accident
- Left or not
- Promotion last 5 years
- Department
- Salary

I mean that an analysis of this dataset will be sufficient in the sense that I do not want to relate the analysis to a specific company (which would require a more specific file), but looking at the problem from a general perspective, and come up with a proposal related to how a company may approach their own equivalent data.

Solution Statement

I will, after cleaning and preprocessing the data, aim at finding the features in the data that correlate the most, by investigating the data in a correlation matrix. Furthermore, I will investigate more on the respective features, using visualizations to explain the correlations. As this is a classification problem, I will also apply supervised classification algorithms like Logistic Regression, Decision Tree Classifier and Random Forest Classifier and see what model makes the most accurate prediction on which employees that will most likely choose to leave.

Benchmark Model

As a reference, I will use base rate, as about 75 % in the dataset have not quit, such an approach would give 75 % accuracy if I said that no one would leave. Of course, I aim to get better prediction than that.

Evaluation Metrics

As the dataset is imbalanced (just about 1 of 4 employees have left), it is appropriate to look at Precision, Recall and F1-score metrics when evaluating the results, rather than the Accuracy. I will also use Cross-Validation, due to the relatively small datasets I get after splitting into train- and test sets.

Project Design

First, the data at hand will be gathered, explored and assessed. Then, the data will be prepared so that machine learning techniques may be applied to it. This preprocessing implies cleaning, formatting and eventual restructuring of the data. Furthermore, any skewness in the data may demand some transformations. Also, scaling numerical features is good practice to perform. Then, the data will be split into test- and train-sets, before applying appropriate models. Further, I will conduct training of the models and making predictions. Finally, I will optimize the model that works best, in order to make conclusions on a recommendable way for companies to get insight on how to keep their employees. The findings will be summarized in a report.