DNB Data Scientist for Enterprise Nanodegree Program

Capstone Proposal

John Gunnar Blostrupsen
November 21st, 2018

Proposal

Domain Background

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.

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 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.

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.

Solution Statement

I will aim at finding the features in the data that correlate the most. I will also look into using classification techniques and test 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

I will use Accuracy score and F-score will be used to evaluate which model that works best in this context.

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.