Machine Learning Engineer Nanodegree

Capstone Proposal

Felipe Tadeu de Souza Lima August 9th, 2018

Proposal

Domain Background

In this project we will assess employee turnover using machine learning classification algorithms. In human resources context, turnover is the act of replacing an employee with a new employee. Partings between organizations and employees may consist of termination, retirement, resignations among other reasons. High turnover may be harmful to a company's productivity if skilled workers are often leaving and the worker population contains a high percentage of novices.

Each company has its own unique turnover drivers so companies must continually work to identify the issues that cause turnover. Further, the causes of attrition vary within a company such that causes for turnover in one department might be very different from the causes of turnover in another department.

Problem Statement

Given a vector of employee data such as Age, distance between work and home, monthly income and others, we'll use classification algorithms to perform a binary classification on whether there's a high probability of a certain employee leaving the company or not. We'll not try to diminish turnover that would be impossible since we're only highlighting employees with a higher tendency to leave the company given a combination of parameters. Our main goal is to evaluate every employee and highlight those which HR department should look into it.

Datasets and Inputs

The dataset used in this project is the HR Employee Attrition and Performance. It's a fictional dataset created by IBM data scientists and contains 1470 rows of employee historical data, each row representing one particular employee. The dataset is open and widely available for data analysis. Some features presented in the dataset are "Age", "Distance from home", "Overtime", "Education, "Marital Status", "Total Working Years", "Monthly income" and others that will be thoroughly during further exploratory data analysis.

Solution Statement

The solution involves the correct prediction of the 'Attrition' column on the dataset. That column indicate if a given employee left the company or not. The prediction should be a simples 'Yes' or 'No', 'Yes' meaning that the employee indeed left the company.

Benchmark Model

According to the available data, 84% of the data is labeled 'No'. A naive approach would be just write 'No' to every employee and we should be right about 80% of the time. We will set this naive approach as a baseline and 84% accuracy as a benchmark goal to be beaten by the ML models.

Evaluation Metrics

The one most important metric will be Accuracy, defined by the 'Number of correct predictions' / 'Total number of predictions'. Further metrics will be used to assess the performance of tested models like precision, recall and F1 score but our goal will be to beat the 'naive' accuracy benchmark of 84%.

Project Design

Since the dataset is tidy, little data processing will be need. We'll encode categorial features but besides that the main focus will be in using many different classification methods such as Logistic Regression, SVM, Neural Networks among others and assess the performance of each model in terms of accuracy.

In a real deployment scenario, the information would probably made available by the company's HR and/or payroll system and later transformed and stored in a relational database. From that database, data would train the machine learning model on a regular basis (weekly or monthly depending on the company) and this model would be deployed as an API to be consumed on dashboards and spreadsheets.

We'll use scikit-learn and tensorflow with Keras for this task since both frameworks have excellent performance regarding classification tasks. Gridsearch will be used to explore hyper-parameters optimization. Prior to algorithm selection there'll be an extensive EDA on the dataset to be used.

The idea around the final solution is that not only a good model should be the outcome but some insights on which features are strong predictors to employee turnover so model interpretability and a good EDA are both important.
