Machine Learning Engineer

Udacity Capstone Proposal

Bruno Marques

October 20th 2018

Capstone Proposal

Domain Background

Currently a lot of companies suffers with employee's attrition, these turnovers causes business to loose qualified personal and valuable assets. Other problem that attrition rates causes are the knowledge lost, sometimes this can cost years of research or thousands of dollars in development to companies, so it is clearly a problem most business would like to avoid or to reduce.

When considering downtime, recruiting, interviewing, training, and getting to speed expenditures are substantial, according to Porter (2011). An entry-level position can cost between 50 to 100 percent of the employee's wage to the organisation (Porter, 2011).

Problem Statement

The proposed capstone project will assess and examine employee attrition rates of company A (real name was hidden due to confidentiality issues), studying employee records will supply supportive information. Obtaining insights from the data will clarify reasons that make employees to leave and help to provide some answers to the questions the company has, such as:

1. What could be changed in the work environment to change the mind of those who wish to leave?
2. What are the main reasons?
3. Is there a location with an accentuated turnover problem?

Presented data will help top management improve decision-making relating to employee work policies in addition to generating new insights on employees retention. The expected result is a diminishing employee turnover and a greater talent retention, as well as lower expenditures with new employees qualification and training.

Datasets and Inputs

The available data about company A employees has information and historical observations over a 8 years period. It consists of data on employees terminations, for each year it shows employees that continued active and those who left.

Solution Statement

Present the development of the following items:

1. your proposed DS approach for the initiative,
2. the findings of the first week (including at least one predictive model as prototype) and
3. the suggested next steps.

More specifically, build a model to predict which employees will leave the company. Also, suggest how the company could leverage the termination data to reduce unwanted attrition.

The first stage is to establish the viability of such a model, by leveraging descriptive statistics and visualisations as means to extract interesting insights from the provided data before diving into the model. Lastly build the prediction model and measure it with a defined metric, The results of the model can be used by HR department to plan a strategy before the employee sends his/her resignation.

Benchmark Model

Logistic Regression is a model based on linear equations that divides the outputs into two groups, this is a really good model for binary classification problems. As mentioned by Kashyap (2018) Logistic Regression presents the best accuracy for this type of dataset when compared to Lasso, Ridge, Decision Trees and Random Forests methods.

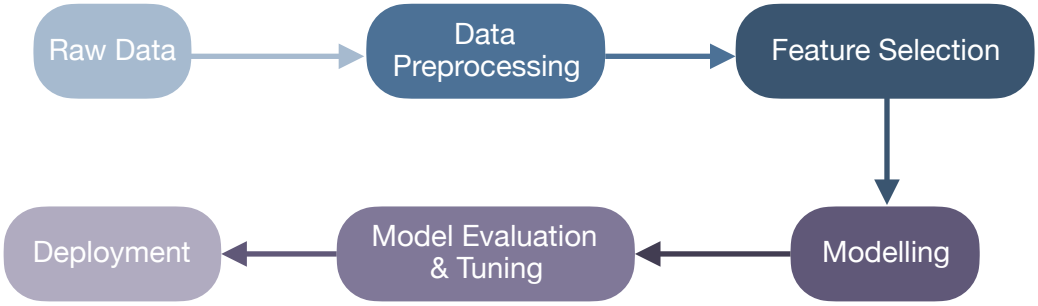
Evaluation Metrics

As we are working with a classification problem, thus the response variable being binary a good metric is the accuracy. Accuracy measures the overall correctness of the model, it sums the total number of true trues and true falses and divide by the total amount of observations. As the employee attrition situation is not something critical it does note need to focus on recall (rate of true trues over the sum of true trues and true falses), neither need to focus on precision (rate of true trues over the sum of true trues and false trues).

Project Design

This project consists of various steps that will help understand, prepare and learn information about the dataset. The main steps are data preprocessing, feature selection & scaling, modelling using Logistic Regression method and finally evaluating the model using the evaluation metric. After evaluation, the model is deployed on the data to make predictions.

Project workflow and descriptions



I. Raw Data

The datasets contain: employee id; employee record date (year of data); birth date; hire date; termination date; age; length of service; city; department; job title; store number; gender; termination reason; termination type; status year; status; business unit.

II. Data Preprocessing

This steps transforms the raw data into a model-able format. Normally real-world data is often incomplete and inconsistent so the following data processing techniques are applied:

- Data Cleaning: Removal of the missing values, smoothening the noisy data or resolving issues with outliers;
- Data Balance: Target features that might present unbalanced observations are balanced through a process of up-sample or down-sample;
- Data Transformation: This process creates new columns, generates more information or consolidate features;
- Data Labelling: This method converts categorical features into numerical, this is required to work with most models;
- Data Reduction: To reduce the dimensions of the data by removing low weighed features.

III. Feature Selection

Feature selection is one of the most important steps during the data analysis. Common methods for this step are correlation analysis, exploratory bivariate analysis and information value analysis. Correlation analysis is used for the numeric variables, as a high correlation proves a feature to be significant.

IV. Modeling

After the data is prepared and model ready it is possible to start the predictive analytics step. The modelling starts by splitting the available data into a training set and a testing set, then the Logistic Regression algorithm is deployed on the training set and tested over the testing set. Randomly 75 % of the data is assigned to the training and 25% to testing. It will be used a classification algorithm since the response variable is binary. The characters in the employee attrition variable are converted into 1 and 0.

V. Model Evaluation & Tuning

Once the Logistic Regression model is trained and fitted it will be evaluated based on accuracy, the score will be saved for latter comparison. Following a hyperparameters search will be done in order to identify the model configuration that will yield the highest accuracy score. Final score is compared with the initial one to assess if there was any improvement.

VI. Deployment

Lastly the model is wrapped-up and deployed to HR department as a minimum viable product (MVP), so it can be used to predict employees which might leave the company in a near future.

References

Porter, J. (2011). Attract and retain top talent. Strategic Finance. 92(12), 56-60. doi:2373925461

Kashyap, Bhuva; Kriti, Srivastava (2018). Comparative Study of the Machine Learning Techniques for Predicting the Employee Attrition. IJRAR July 2018 , Volume 5, Issue 3.

Frye, Alex; Boomhower, Christopher; Smith, Michael; Vitovsky, Lindsay; and Fabricant, Stacey (2018) "Employee Attrition: What Makes an Employee Quit?," SMU Data Science Review: Vol. 1 : No. 1 , Article 9.