Features Engineering for Workforce Analytics

Lily Chan
CSCI 4502
Fall 2022
University of Colorado at Boulder

Michael Ghattas
CSCI 4502
Fall 2022
University of Colorado at Boulder

Khoi Phan
CSCI 4502
Fall 2022
University of Colorado at Boulder

Introduction

Data science is a fast-growing field dominating the decision-making process within the business world and supported by the need for operational efficiency and Return On Investment (ROI) growth mindset. The driving force behind such initiatives is usually the workforce intelligence teams within the Human Resources department.

The project aims to develop a methodology for a supervised ML algorithm utilizing a well-defined series of techniques. The goal is to classify, transform and interpret the aggregated data, and extract the datasets valuable as predictors for a workforce analytics model to measure candidate/employee metrics.

Proposed Work

Our project will tackle the challenges of data aggregation, data cleaning and scrubbing, and data transformation from multiple datasets. Once the data transformation into a uniform dataset is complete, we will proceed to complete an Exploratory Data Analysis (EDA), testing relationships between all the dimensions and merging or reducing those with identified relationships. At this stage, the final dataset is complete, and thus we will begin testing different models to identify the significant predictors of positive outcomes. We will begin with several different datasets from various sources. Some datasets are provided by organizations like IBM, while others are autogenerated by their respective creators.

Evaluation

Given our scope and its relation to the documented inherent bias in the social sciences, we will attempt to understand the data's behavior and distributions to interpret our observations. Additionally, we will document our reasoning and logic for each chosen direction to provide the needed justification for our methods and their applications. Finally, we will assess our interpretation against our results from a linear regression model, identifying each predictor's significance to the desired response, correlation to other predictors, and other notable relationships.

Related Work

Determinants of employee engagement and their impact on employee performance

https://www.emerald.com/insight/content/doi/10.1108/IJPPM-01-2013-0008/full/html

The study collected data from 393 questionnaire responses. The study-targets were middle to low managerial level employees of small-scale organizations. The study aims to identify what factors can influence employee engagement. Factors studied were workplace wellbeing, compensation, coworker relations, leadership style, work environment and policies, and training. The results of regression using all factors resulted in an r^2 value of 0.672, which shows that these factors significantly impact employee engagement.

Factors affecting employee performance of PT.Kiyokuni Indonesia

https://www.emerald.com/insight/content/doi/10.1108/IJLMA-03-2016-0031/full/html

This study was similar to the one that collected data from 451 questionnaire responses. The data was gathered from workers from an Indonesian textile company. The goal of this study was similar to determine the impact that a select number of factors had on employee performance. These factors were leadership, motivation, and discipline. Regression analysis found that all regression coefficients of the factors were positive, implying that all factors improve employee performance, with leadership having the most negligible impact.

Factors affecting employee performance: an empirical approach

https://www.emerald.com/insight/content/doi/10.1108/IJPPM-01-2018-0012/full/html#sec003

The study aims to find relationships between environmental, job, and employee factors by creating an assessment model. Data were collected from questionnaires given to employees and HR directors from various participating firms resulting in a valid sample of 79 HR managers and 392 employees. The model determined that all included factors impacted employee performance, with job environment and adaptability having the most significant impacts and dynamic environments and management support having the least.

An Effectiveness of Human Resource Management Practices on Employee Retention in Institute of Higher learning

https://www.cscjournals.org/manuscript/Journals/IJBRM/Volume3/Issue2/IJBRM-81.pdf

The study sought to find out what factors contribute to employee retention. This study involved data from 278 university faculty who were given questionnaires. Linear regression was done on the data to get an r^2 value of 0.680, concluding that employee empowerment, compensation, and training significantly impact employee retention. The limitations of this study are that all data was collected from the same university in Malaysia, which resulted in a limited sample. In addition, work ethic and culture differences may cause discrepancies compared to data from the U.S.

Milestones

Project Development

Idea generation & conceptualization

Subject research

Data research & collection

Write & review our proposal

Project Implementation

Develop & finalize our plan

Data exploration & assessment

Data aggregation & transformation

Data analysis & interpretation

Project Delivery

Document our results & findings

Write & review our report

Design & review our presentation

Rehearse team presentation

Methods

Languages

R and Python

Environments

JupyterLab and VSCode

Algorithms & Models

Dynamic Discretization

Clustering

Deviation

Imbalance Ratio

Hypothesis Testing

Akaike information criterion (AIC)

Bayesian information criterion (BIC)

K - Nearest Neighbor (KNN)

Monte-Carlo Markov Chains (MCMC)

Data

Absenteeism

https://www.kaggle.com/code/hypnobear/absenteeism-at-work-dataset/data

Size: 65KB

Number of files: 2

File Type(s): CSV and excel

Dimensions: 742 Rows x 21 Columns

Human Resources

https://www.kaggle.com/datasets/rhuebner/human-resources-data-set

Size: 223KB

Number of files: 1

File Type(s): CSV

Dimensions: 1471 Rows x 35 Columns

Turnover

https://www.kaggle.com/datasets/davinwijaya/employee-turnover

Size: 82KB

Number of files: 1

File Type(s): CSV

Dimensions: 1130 Rows x 16 Columns

Job Classification

https://www.aihr.com/blog/hr-data-sets-people-analytics/

Size: 5KB

Number of files: 1

File Type(s): CSV

Dimensions: 67 Rows x 14 Columns

IBM-HR

https://www.aihr.com/blog/hr-data-sets-people-analytics/

Size: 223KB

Number of files: 1

File Type(s): CSV

Dimensions: 1471 Rows x 35 Columns

End.