Potential Performance Predictor - P^3
CSC 522 - Group 19

Subodh Gujar
sgujar@ncsu.edu

Deep Mehta
dmmehta2@ncsu.edu

Kartiki Bhandakkar
kbhanda3@ncsu.edu

1 PROJECT IDEA

The aim of this project is to build a machine learning model/s that can predict employee performance during hiring process. The model will consider existing employees in the company and utilizing their performance over the last years, evaluate how a new employee with similar attributes will fit into the work environment at that company. This approach is unique in the sense that each person responds differently to various aspects of a work environment, and if we utilize the fact that an existing employee in that particular role has been performing well in that work environment, then it would be more accurate rather than just comparing skills or education. A major and ethical issue that we will be addressing as part of this project is to look into attributes that may lead to bias and make decisions to remove any sort of bias during the hiring process. This project will involve Data Analysis, Feature Selection, Model Training, Model Evaluation and ML-Ops. This project will also involve creation of a ML pipeline on Azure, that will automatically train and learn as new data comes in, and also provide APIs to make predictions during the hiring process.

As a part of the novelty aspect in this project, we will perform exploratory data analysis of dataset to understand feature co-relation and experiment with different features to identify bias in the model and remove it. Also, we will be combining different machine learning models, such as through ensemble methods, to increase accuracy and improve performance of the overall model. Considering the use case, this dataset will also be used to draw meaningful hypothesis, that goes beyond simple 'yes' and 'no'.

2 DATASET

We are considering the following dataset for our project:
IBM HR Analytics Employee Attrition & Performance (Link).

This is a fictional data set created by IBM data scientists. This dataset describes 35 attributes related to 1500 employees across the company.

This dataset provides relatable attributes like Department, Education, Environment Satisfaction, and many more, that we believe will be a good starting point for us to train the machine learning model and make suitable predictions on performance of employee as described in the project idea above.

3 SOFTWARE

In terms of software we will be utilizing Jupyter notebooks to write python scripts for training, testing and experimentation. In terms of libraries, we will use common libraries like Scikit-Learn, Pandas, Numpy etc. This is not an exhaustive list, and we may use other suitable libraries as we learn about them from research and experimentation. In addition to this, we also intend to use Azure, to create an automated machine learning pipeline (ML-Ops).

4 RESEARCH PAPERS

[1] A. A. Mahmoud, T. AL Shawabkeh, W. A. Salameh and I. Al Amro, "Performance Predicting in Hiring Process and Performance Appraisals Using Machine Learning," 2019 10th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 2019, pp. 110-115, doi: 10.1109/IACS.2019.8809154. (Link)

[2] Al-Radaideh, Qasem and Alnagi, Eman. (2012). Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance. International Journal of Advanced Computer Science and Applications.3.10.14569/IJACSA.2012.030225.(Link)

[3] Mosavi, A., Sajedi Hosseini, F., Choubin, B. et al. Ensemble Boosting and Bagging Based Machine Learning Models for Groundwater Potential Prediction. Water Resour Manage 35, 23–37 (2021). https://doi.org/10.1007/s11269-020-02704-3 (Link)

[4] Kotsiantis, Sotiris & Pintelas, P.. (2005). Combining Bagging and Boosting. International Journal of Computational Intelligence. 1. 324-333(Link)

5 WORK DIVISION

The work will be equally distributed among all the three team members as described below:

(1) Subodh Gujar: Evaluate supervised ML techniques, Experiment and Analyse Results
(2) Deep Mehta: Evaluate unsupervised ML techniques, Experiment and Analyse Results
(3) Kartiki Bhandakkar: Data Pre-processing, Feature selection, Feature creation

In the end, results will be compared using various techniques and their combinations, and utilize that to decide the next steps.

6 MIDTERM MILESTONE

In terms of Midterm Milestones, we plan to complete following tasks:

(1) Research techniques that have been used in the past for similar use case.
(2) Experimentation with multiple machine learning techniques and analyze the results.
(3) Feature selection, reduce attributes and reduce model complexity, while ensuring high model accuracy along with reduction in bias.
(4) Train model with at least two different machine learning techniques.
(5) Fine tuning the hyper parameters to improve the performance and accuracy of model.