Literature Review of Machine Learning Models for Screening Job Applicants

---

Introduction

The process of screening job applicants plays a crucial role in identifying the most suitable candidates for interviews. With the increasing volume of applications, manual screening becomes labor-intensive and time-consuming. To address this challenge, researchers have turned to machine learning models to automate the screening process based on applicants' resumes and cover letters in fact Forbes [i]reports that most applications processes go through an internal database that selects applicants based on matches with key qualifications and skills.

Machine learning models provide the following advantages over conventional screening processes[ii].

- Efficiency and Time Saving (processes information faster)
- Objective and consistent Evaluation (processes without human bias and error)
- Improved Accuracy and Predictive Capabilities (Improves accuracy of potential candidates.
- Scalability (can continue to process and handle large sets of data)
- Automated Feature Extraction (can pull important information faster)
- Continuous development and tuning (can be improved with every new set of data)

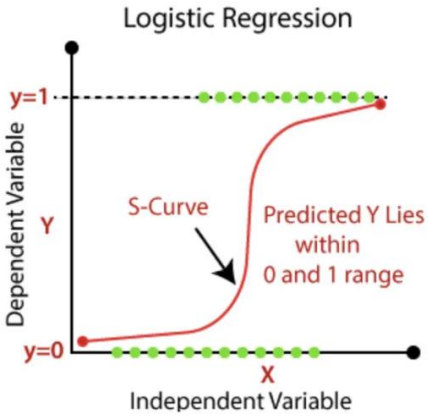
This literature review will examine three machine learning models and determine which is most optimal for screening job applicants for interview, given 100+ applicants resumes and cover letters, the model will rank them 0 -100 and only select the top 10 candidates. In order to determine suitability and which is the most optimal we will need to examine key metric such as accuracy, precision, recall and interpretability.

Machine Learning Model 1:

Logistic Regression:

Logistic regression is a commonly used model for binary classification tasks. In the context of screening job applicants, it can be adapted to assign a binary label (e.g., shortlist or reject) to each candidate. This model can be used to estimate the probability of an applicant being selected for an interview based on their resume and cover letter contents. This can allow the screening

process to key in on key words, or experiences and assign a weight rating on them and short list candidates based on their weight rating[iii].



Pros:

- Simplicity and ease of interpretation.
- Fast training and prediction times.
- Works well with small to medium-sized datasets.

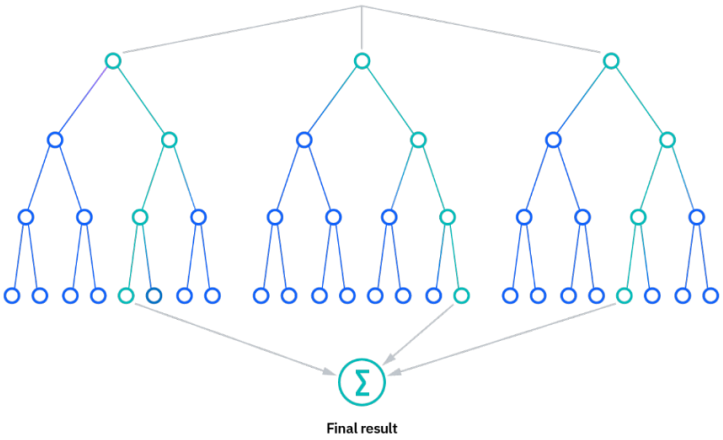
Cons:

- Assumes a linear relationship between features and scores.
- May not capture complex relationships in the data.
- Performance might be affected by outliers or noise in the data.

Machine Learning Model 2:

Random Forest:

Random Forest is an ensemble model that combines multiple decision trees to make predictions. It has gained popularity in applicant screening due to its ability to handle high-dimensional data effectively and capture complex relationships. Random Forest models are less prone to overfitting compared to individual decision trees. However, the interpretability of Random Forest models is lower than that of logistic regression. Training and prediction times are slower compared to logistic regression, and hyperparameter tuning is necessary to optimize performance[iv].



Pros:

- Ability to capture complex relationships between features and scores.
- Robust against outliers and noise.
- Handles high-dimensional data effectively.

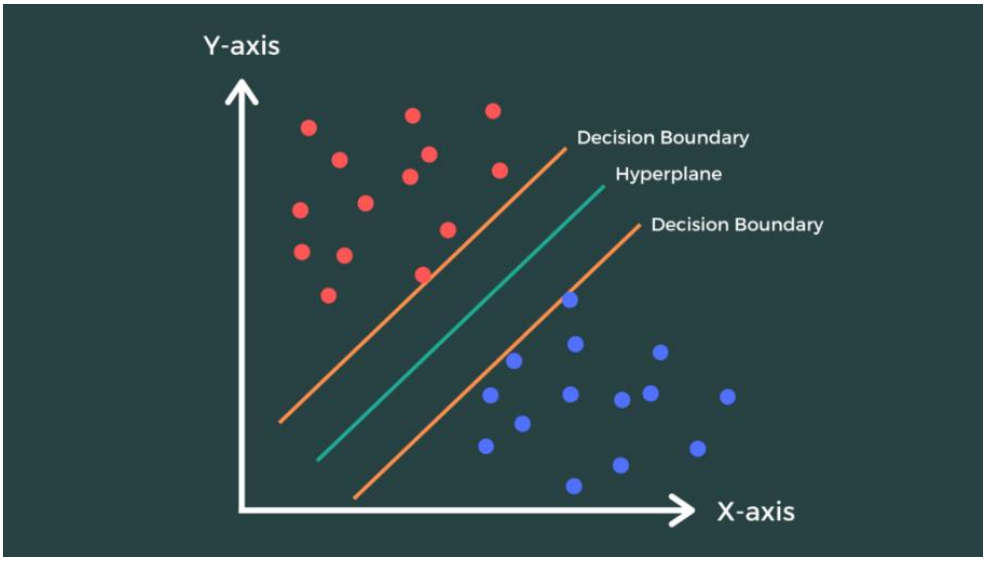
Cons:

- Less interpretable compared to linear regression.
- Slower training and prediction times compared to linear regression.
- Requires tuning of hyperparameters.

Machine Learning Model 3:

Support Vector Machine (SVM):

Support Vector Machine is a powerful and versatile machine learning model that can handle both linear and non-linear classification tasks. SVM aims to find an optimal hyperplane that separates different classes with the maximum margin. It can handle high-dimensional feature spaces effectively through the use of kernel functions, which transform the input features into higher-dimensional spaces. SVM has been widely used in various domains, including applicant screening[v].



Pros:

- Effective in handling high-dimensional feature spaces.
- Ability to handle non-linear relationships through kernel functions.
- Provides a clear margin of separation between different classes.
- Can handle both binary and multi-class classification tasks.
- Generally good performance even with limited training data.

Cons:

- Choosing an appropriate kernel function and tuning hyperparameters can be challenging.
- Interpretability is often lower compared to linear regression.
- Computationally more expensive than linear regression.
- Does not naturally handle imbalanced datasets, requiring additional techniques like class weighting or resampling.

Reference:

[i] Ryan, Liz. “How Can 100 Job Applications Get Zero Replies? Here’s How.” *Forbes*, 28 Apr. 2017, www.forbes.com/sites/lizryan/2017/04/28/how-can-100-job-applications-get-no-replies-heres-how/?sh=14265f871e55.

[ii] Lokesh. S, et al. “Resume Screening and Recommendation System Using Machine Learning Approaches.” *Computer Science & Engineering: An International Journal*, vol. 12, no. 1, 28 Feb. 2022, pp. 1–7, https://doi.org/10.5121/cseij.2022.12101. Accessed 27 July 2022.

[iii] Kumar, Narender. “Logistic Regression Explained with Examples.” *Spark by {Examples}*, 12 Mar. 2023, sparkbyexamples.com/machine-learning/logistic-regression-explained-with-examples/.

[iv] IBM. “What Is Random Forest? | IBM.” *Www.ibm.com*, www.ibm.com/topics/random-forest.

[v] “Support Vector Machine (SVM) Classifier - the Click Reader.” *Www.theclickreader.com*, 24 Nov. 2020, www.theclickreader.com/support-vector-machine-svm-classifier/. Accessed 21 May 2023.