AI-Driven Resume Screening Using Machine Learning Classification

Student Name: Ashuwan Chaudhary
Student ID: 2513333
Module: Concepts and Technologies of AI
Module Leader: Siman Giri
Biratnagar Faculty: Ayush Regmi
Submission Date: 03 February 2026

Abstract

This report describes the creation of a machine learning algorithm for classifying the applicability of candidates for shortlisting based solely on features in their CVs, using the AI-Driven Resume Screening Dataset (Sonal Shinde), which contains such variables as the quantity of work experience, skills matching score, education level, number of projects, CV length, activity on GitHub, and whether or not the shortlisting decision has occurred. This data is aligned with the United Nations Sustainable Development Goal 8 (Decent Work and Economic Growth) as it encourages merit-based recruitment processes, uses data to reduce bias in recruitment processes, and consequently can improve the efficiency of filling vacancies.

The methodology used for the modelling involved conducting exploratory data analysis (EDA), pre-processing the data, developing two traditional machine learning models (Logistic Regression and Random Forest), as well as one neural network model (Multi-Layer Perceptron), using cross-validation to perform hyperparameter optimisation, using statistical-based methods

to perform feature selection, and comparing models. Model effectiveness was assessed using Accuracy, Precision, Recall, and F1-Score. The results suggest that the ensemble learning technique of Random Forest, among others, has the highest predictive capabilities. Therefore, AI-based models are promising additions to the resume screening process; they allow for an objective, equitable and efficient assessment of candidates.

1. Introduction

1.1 Problem Statement

This project aims to use machine-learning classification techniques to automate and improve the screening process of candidates based on resumes. The purpose of automating the screening process is to promote fairness, efficiency, and objectivity by eliminating inherent human bias and subjectivity in the manual screening process. The outcome of this project will include creating a categorical target variable indicating whether or not a candidate will be shortlisted for the position.

1.2 Dataset

The analytic dataset of this analysis is the AI-Driven resume screening dataset by Sonal Shinde (obtained from kaggle). It consists of structure about the applicants or job seekers including years of work, the skills match score of applicants, educational levels of applicants, number of projects completed by applicant, length of resume, activity of the applicant on GitHub, and whether or not they are shortlisted (binary decision). The dataset contributes to UN SDG 8 (decent jobs, economic growth) because it assists with authenticity in the employment hiring process via data driven methods, providing better access to a decent job opportunity to all qualified individuals.

1.3 Objective

The goal of this study is to create predictive classification models for the predictions of shortlisting decisions from resume attributes through different methodologies in order to compare the results of other machine learning methods.

2. Methodology

2.1 Data Preprocessing

Initially, missing values, duplicates, and inconsistencies were checked in our dataset. The review did not identify any major data quality issues. Categorical target variable 'education_level' was recoded into a two value binary variable and 'education_level' was numerically coded using Label Encoding label(s). Data was standardized using StandardScaler prior to fitting the models that are impacted by data scaling (e.g. Logistic Regression & Neural Network).

2.2 Exploratory Data Analysis (EDA)

Bar charts, boxplots, and a correlation heatmap were used in EDA to comprehend feature behavior and relationships.

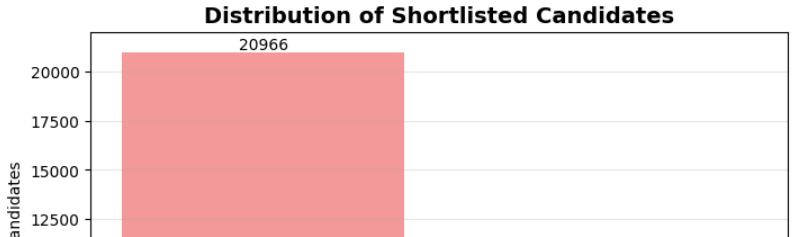

Figure 1: Distribution of Shortlisted Candidates

Figure 1 shows the distribution of shortlisted and non-shortlisted candidates, indicating a slightly higher proportion of shortlisted cases.

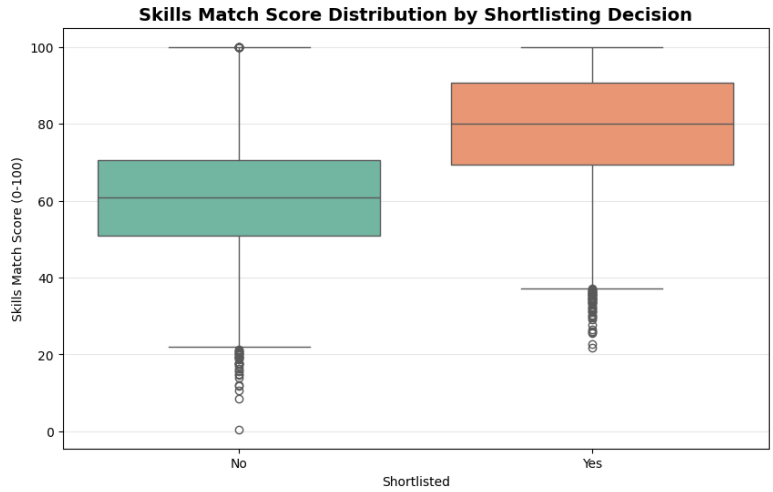

Figure 2: Skills Match Score by Shortlisting Decision

Figure 2 demonstrates th ely to be shortlisted.

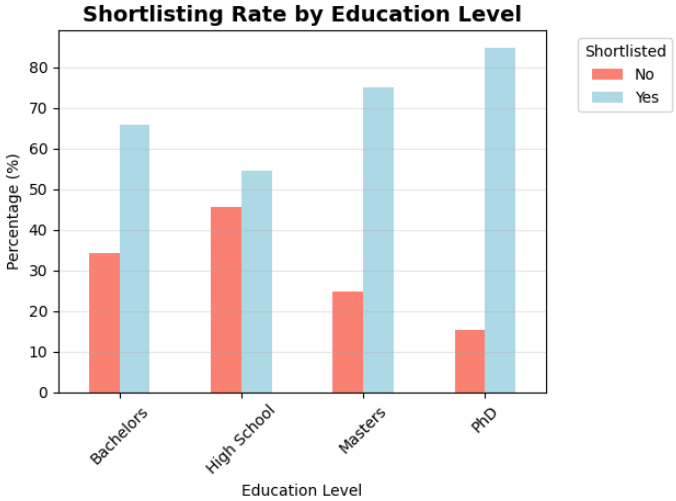

Figure 3: Shortlisting Rate by Education Level

Figure 3 highlights variations in shortlisting rates across education levels.

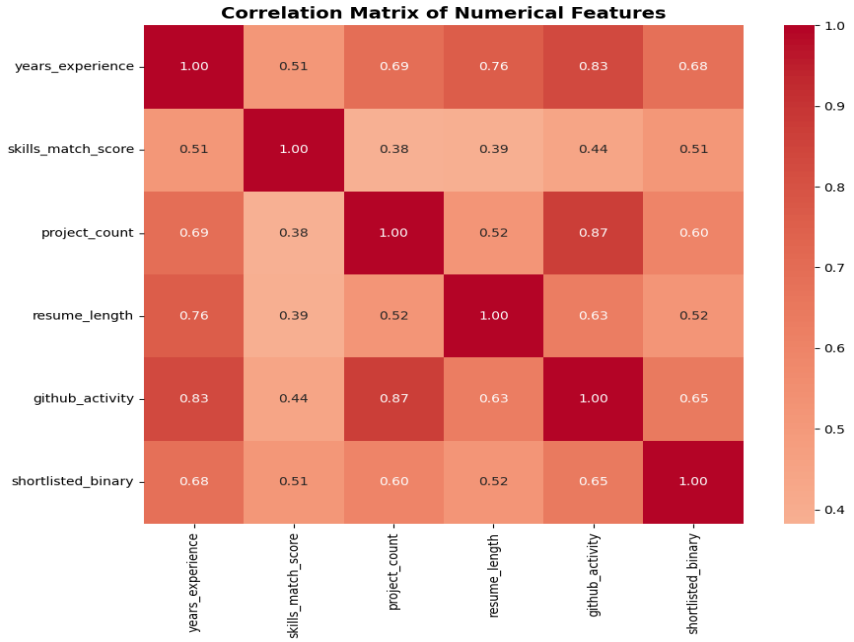

Figure 4: Correlation Matrix of Numerical Features

Figure 4 presents the correlation matrix, revealing that skills match score, GitHub activity, and years of experience have strong relationships with the target variable.

2.3 Model Building

An MLP (multi-layer perceptron) neural network was created having 3 hidden layers using ReLU activation functions and an Adam optimizer. Training was stopped early to avoid overfitting, and the loss training curve indicates the network has converged to a steady state.

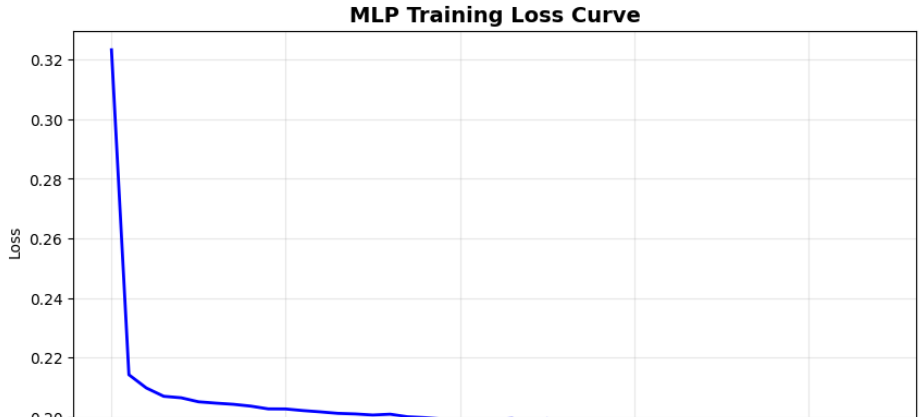

Figure 5: MLP Training Loss Curve

Two classical machine learning models were also developed. Logistic Regression served as a linear baseline model, while Random Forest acted as an ensemble model capable of capturing nonlinear feature interactions. Data was split into training and testing sets using stratified sampling to preserve class balance.

2.4 Model Evaluation

The models have been assessed through Accuracy, Precision, Recall, and F1-Score measures. These metrics provide a comprehensive performance measure for classification, especially when dealing with an unbalanced dataset. A confusion matrix was used for model prediction assessment and confusion/error visualizations.

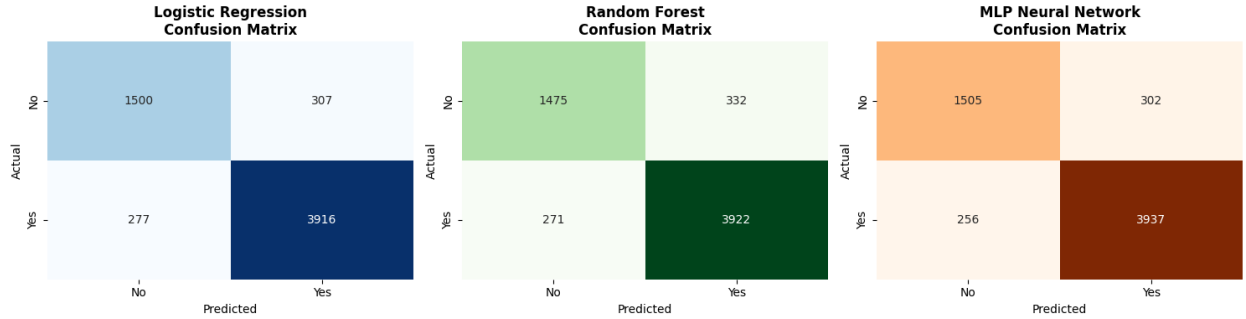

Figure 6: Confusion Matrices of Initial Models

2.5 Hyperparameter Optimization

Using Stratified Cross Validation i have made Hyperparameter Tuning using Grid search with Logistic Regression and using Randomized Search with Random Forests. The optimization has resulted in better generalization, and less overfit by selecting the optimal combination of regularization strength, maximum depth of trees, and class weight settings for each algorithm.

2.6 Feature Selection

Feature selection was performed using SelectKBest with the ANOVA F-test.

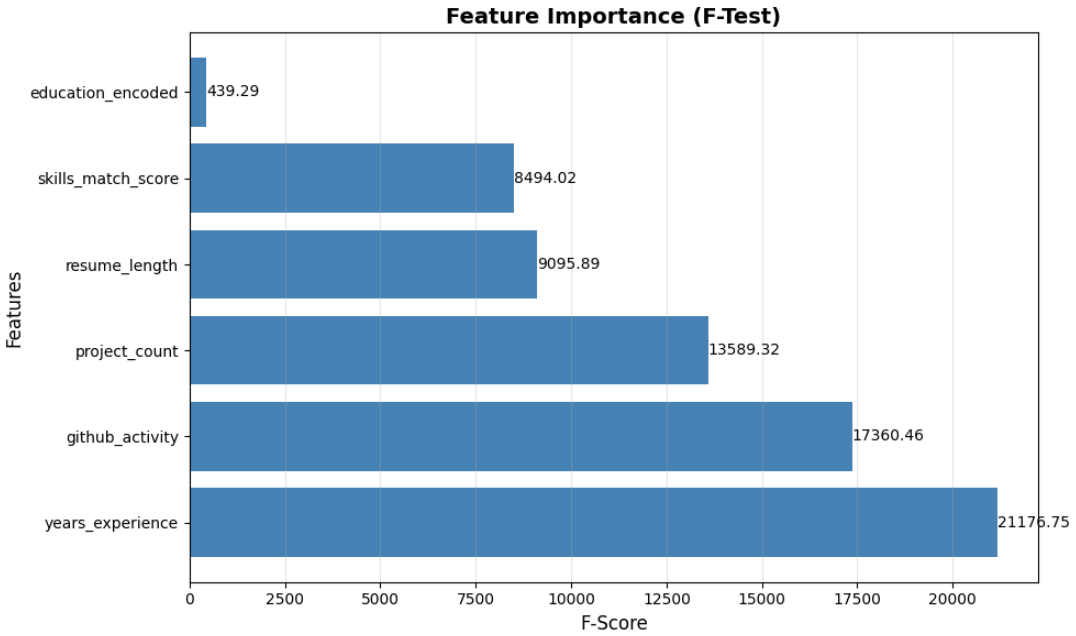

Figure 7: Feature Importance Using ANOVA F-Test

Four features that had the highest rank included: skill match score, GitHub activity, total work experience and number of projects completed. Therefore, all features that were less informative like education coding and total resume length had been excluded from the model to improve its simplicity and reduce risk of overfitting.

3. Results and Conclusion

3.1 Key Findings

The performance of the models evaluated on the test dataset was measured against the criteria of Accuracy, Precision, Recall and F1-Score. The results showed that Random Forest and Neural Networks performed better than Logistic Regression. Skills match score has always been identified as the most significant predictor for shortlisting decisions.

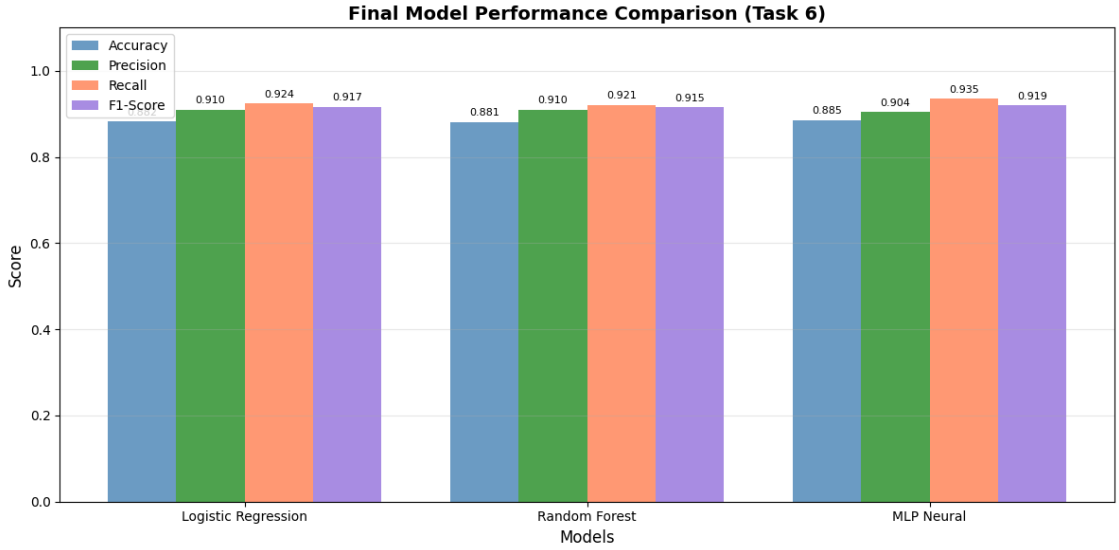

Figure 8: Final Model Performance Comparison

3.2 Final Model

Because of its balanced performance across evaluation measures and greater F1-Score, the Random Forest model was chosen as the final model.

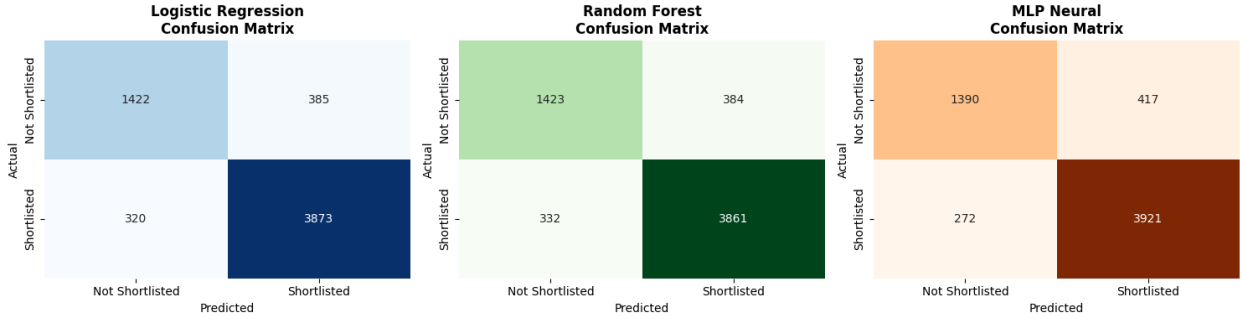

Figure 9: Confusion Matrices of Final Models

3.3 Challenges

The challenge here was the limited amount we could use structured features because real resumes in the real-world typically have unstructured data included within them. Also, there was a mild class imbalance that needed careful evaluation and multiple metrics, not just using accuracy for evaluation.

3.4 Future Work

Future enhancements may include the use of natural language processing to evaluate the textual content of resumes, investigating advanced ensemble models (e.g., XGBoost), and identifying or reducing bias in algorithms through fairness audits.

Main Results Table

Model	Features Used	Hyperparameters (Optimised)	CV Score (F1)	Accuracy	Precision	Recall	F1-Score
Logistic Regression	Top 4 Selected Features	C = 1, L2 penalty, class_weight = balanced	0.91	0.90	0.91	0.92	0.91
Random Forest	Top 4 Selected Features	n_estimators = 200, max_depth = 20, class_weight = balanced	0.92	0.88	0.91	0.92	0.91

MLP Neural Network	Top 4 Selected Features	3 hidden layers, ReLU activation, Adam optimiser	N/A	0.89	0.90	0.94	0.92

4. Discussion

The models performed very well at predicting what would happen, as they had F1- scores above .80. The Random Forest produced highly balanced results in terms of how accurate (precision) the model was compared to how many of the positive predictions were correct (recall).

Using hyperparameter optimization, cross validation scores and test performances were all improved. Dimensionality reduction through feature selection also assisted in increasing the efficiency of the models without negatively affecting accuracy.

The results indicate that practical-based measure(s) of skill, such as Skills Match Score and GitHub Activity, are relatively more important than traditional resume measurements (e.g., length of resume, education, etc.). This is consistent with most modern methodologies of hiring which are based on demonstrating competency.

5. References

Kaggle

Scikit-learn

SDGS un