OPTIMIZING CANDIDATE SELECTION USING RECRUITMENT DATA

FINAL PROJECT REPORT

Presented By

HECTOR SANCHEZ

Introduction

Problem Identification:

- Hiring and Operations managers can often find it difficult to find high quality candidates to join their teams. Recruiting top talent requires a lot of energy, time, and attention to detail. In many cases, hiring decisions are made out of desperation due to staffing issues which can often have adverse effects on the performance of businesses on the front line, regardless of the industry. These hiring and operations managers would benefit greatly from having a predictive model that can determine which candidates are most likely to be hired based on their profiles. This would allow leaders to focus their efforts on finding high-potential candidates during the recruitment process.

Objective

- The goal of this project is to build a predictive model that can accurately classify whether a candidate is likely to be hired, using data from the recruitment data dataset.
- This project ventures into building different models, applying hyperparameter tuning to optimize these models, and it covers the process of selecting the best performing model based on evaluation metrics

Data Overview

Dataset Description

- Predicting Hiring Decisions in Recruitment Data
- Author: Rabie El Kharoua
- This dataset is synthetic, original, meant for educational purposes, and is owned by the author listed above.
- The dataset that is used in this project is the Recruitment Data Dataset, which contains features such as age, gender, education level, experience years, and many others. The target variable in this case is Hiring Decision since it has a binary nature (a candidate is either, 1 = hired, or 0 = not hired

Key features include:

- Demographic: **Age, Gender, DistanceFromCompany,**
- Qualifications: **EducationLevel, SkillScore, PersonalityScore, InterviewScore**
- Experience: **ExperienceYears, PreviousCompaniesWorked**
- Target: **HiringDecision**

Data Overview Cont’d

Here is a breakdown of every feature in the dataset:

- **Age** (integer): Candidate's age.
- **Gender** (binary): Male (0) or Female (1).
- **EducationLevel** (ordinal): Education level ranging from 0 (No Formal Education) to 4 (Postgraduate).
- **ExperienceYears** (integer): Years of work experience.
- **PreviousCompanies** (integer): Number of companies a candidate worked at previously.
- **DistanceFromCompany** (float): Distance (in miles) from the company.
- **InterviewScore, SkillScore, PersonalityScore** (0-100): Evaluation scores.
- **RecruitmentStrategy** (categorical): Recruitment strategy applied (0 = Undefined, 1 = Aggressive, 2 = Moderate, 3 = Conservative).
- **HiringDecision** (binary): Whether the candidate was hired (0 = No, 1 = Yes).

Data Wrangling

Missing Values

- This particular dataset did not have any missing values, so we did not need to drop any values or use imputation.

Feature Mapping

- We applied feature mapping in order to make the ‘Gender’ and ‘EducationLevel’ features easier to interpret
- Gender: 0 mapped to ‘Male’, and 1 mapped to ‘Female’
- EducationLevel: 1 mapped to “Bachelor’s Type 1”, 2 mapped to “Bachelor’s Type 2”,, 3 mapped to “Master’s,”, and 4 mapped to “PhD

One-Hot Encoding

- We applied One-Hot Encoding to “RecruitmentStrategy” and applied the following suffixes:
    - Strategy_Aggresive = 1
    - Strategy_Moderate = 2
    - Strategy_Conservative = 3

Feature Scaling

- Numerical features such as “Age” and “InterviewScore” were standardized in order to ensure that the models are trained on similar feature scales.

Exploratory Data Analysis (EDA)

Key Insights:

- **Correlation Analysis:** The EDA that was performed makes it clear that the features, **Strategy_Aggresive, InterviewScore, PersonalityScore,** and **SkillScore**, all have a clear, positive correlation with **HiringDecision** (our target feature).

- Candidates that were sought out via an Aggressive recruitment strategy demonstrated the highest hiring rate. This suggests that employers were more likely to hire a candidate if their approach to recruiting was more intentional (aggressive). This can imply many things, but it's safe to say that recruiters/managers that use an Aggressive strategy have a clear, urgent need to hire, which is what likely results in the higher hiring rate.

- **DistanceFromCompany** demonstrates a weak and negative correlation with **HiringDecision**. Therefore, a candidate's likelihood of being hired was hardly impacted by how long their daily commute would be.

Visualizations

HIstograms for Age, ExperienceYears, PreviousCompanies, DistanceFromCompany, InterviewScore, SkillScore, and PersonalityScore

- The first set of visualizations that I used was a series of histograms. Each histogram displays the distribution of the numeric features from the dataset.

- **Distribution of Age:** This histogram shows the spread of ages for the candidates contained in this dataset. The candidates' ages seem to be evenly distributed, which suggests that candidates of all ages (within the range found in the dataset) were seen as hireable.

- **DIstribution of ExperienceYears:** This histogram shows the range and frequency of a candidates' work experience in years. The distribution appears to be quite wide, which suggest that the dataset contains candidates that range from being entry-level to being highly experienced professionals.

- **Distribution of PreviousCompanies:** This distribution displays the number of companies that a candidate has worked for previously. Candidates who have worked for many companies could arguably contain a higher amount of adaptability, while those with longer tenures at less companies show signs of loyalty and stability.

Visualizations Cont’d

- **Distribution of DistanceFromCompany:** This histogram displays how far candidates live from the company’s location. The distribution seems to be evenly distributed, so there are many candidates that live close, far, and in between in relation to the company’s location. However, it’s worth noting that candidates that live further away could possibly show lower signs of retention due to their longer commutes.

- **Distribution of InterviewScore:** This histogram shows each candidate’s performance in an interview, placed on a scale of 1-100. Candidates with higher interview scores could have a higher likelihood of being hired.

- **Distribution of SkillScore:** This distribution shows the range of technical skills that a candidate has. Higher SkillScores could suggest that a company values skilled candidates if it leads to more job offers.

- **Distribution of PersonalityScore:** This histogram displays a candidate’s interpersonal or behavioral attributes on a scale of 0-100. Candidate’s that score higher with this metric could have a higher likelihood of fitting in with the company and team culture.

Visualizations Cont’d

Bar Plots for Categorical Feature Analysis

- In order to assess the distribution of the dataset’s categorical features, we used a series of Bar Plots.

- **Distribution of Gender**: This bar plot displays the count or proportion of male (0) and female (1) candidates that are in the dataset. Since the distribution is more balanced, this suggests that the dataset reflects an equal amount of representation between Males and Females in the hiring process. Having a balanced gender pool helps us understand if hiring decisions are influenced by gender or not.

- **Distribution of EducationLevel:** This bar plot shows the count of candidates with respect to their level of education. In this dataset, we have 4 levels of education. If most candidates that are hired are from a specific education level, a company can focus on attracting those candidates in order to improve their hiring efforts.

Visualizations Cont’d

- **Distribution of Strategy_Aggressive:** This bar plot displays the number of candidates that were recruited via an aggressive recruitment strategy.

- **Distribution of Strategy_Moderate:** This bar plot shows the number of candidates that were recruited via a moderate recruitment strategy.

- **Distribution of Strategy_Conservative:** This bar plot displays the count of candidates that were recruited via a conservative recruitment strategy. This strategy is more effective for filling senior level positions where companies may need to be more selective and careful when extending offers.

Visualizations Cont’d

Correlation Heatmap for Correlation Analysis

- We computed a correlation matrix so that we could plot a correlation heatmap. The values that are shown are the Pearson correlation coefficients, which gives us insight into the linear relationship between pairs of features.
  - The heatmap indicates the strength and direction of feature correlations via the use of color gradients. Darker, more concentrated colors represent a higher correlation coefficient. In this case, Dark Blue represents a high negative correlation, while Dark Red represents a high positive correlation.
  - This visualization helps us reveal which features have a strong relationship with the HiringDecision feature.
  - Based on the heatmap, the following correlations/relationships with respect to HiringDecision are apparent:
    - **Positively Correlated:** Strategy_Aggressive
    - **Slightly Positively Correlated:** EducationLevel_Master’s, EducationLevel_PhD, InterviewScore, SkillScore, PersonalityScore, and ExperienceYears
    - **Negatively Correlated:** Strategy_Moderate, Strategy_Conservative, EducationLevel_Bachelor’s, Type 2

Visualizations Cont’d

Box Plots and Cross Tabulation to Inspect Feature Relationships

- We used Box Plots to inspect the relationships between HiringDecision and the numerical features.
    - Based on these Box Plots, it seems that Interview/Personality/Skill Score have the most visible impact on someone being hired or not.

- We used unstacked Bar plots to examine how each of our categorical columns related to HiringDecision. We observed the following based on the Bar Plots.

    - The Gender gap between Male and Female candidates is not incredibly significant, but it does seem that more Males are hired versus Females.

    - Candidates with an EducationLevel of Bachelor’s Type 1 and PhD were hired the least. These candidates could be seen as either being under qualified or overqualified.

Summary of Findings from Visualizations

- Scoring high in interview and personality metrics plays a critical role in hiring decisions. Recruiters should place an emphasis on evaluating these metrics effectively.

- Aggressive recruitment strategies demonstrate the highest correlation with successful hiring decisions, which suggests that it should be prioritized for larger talent acquisition efforts. Conservative strategies can be considered when searching for senior level, niche candidates.

- The balanced representation of gender and age suggests that we have inclusivity in the hiring process (in the context of this dataset). The slight imbalance suggests that this be monitored further if more data is acquired down the line.

- Retention can be improved if organizations consider offering flexible work arrangements or relocation support for candidates that live further from the company.

Modeling Process

Data Preparation

- We performed a train/test split where the dataset was split into 80% training and 20% testing

- The split resulted in X_train having 1200 rows and 14 columns, and X_test having 300 rows and 14 columns

Models Implemented

- Logistic Regression: Effective and simple Baseline model that is often used for binary classification tasks. In this case, it’s helping us map predicted values between 0 (not hired) and 1 (hired).

- Random Forest Classifier: Versatile model that handles both categorical and numerical features while performing well on imbalance datasets

- Support Vector Classifier: Flexible model that reduces overfitting by maximizing the margin between classes.

Modeling Process Cont’d

H**yperparameter Tuning**

- Optimized Random Forest performance with **GridSearchCV**. This technique trains the model using all combinations of the hyperparameters across multiple folds in a cross validation process.
    - GridSearchCV helped explore the combinations of parameters like n_estimators, max_depth, and min_samples_split, which resulted ina model with the best balance of underfitting and overfitting.

- Optimized Support Vector Classifier performance with **RandomizedSearchCV**. This technique randomly selects a defined number of combinations from the distributions. Each selected combination is evaluated using cross validation to measure performance.
    - RandomSearchCV optimized parameters like C and gamma, which allowed for better performance while saving time compared to a full grid search

Model Evaluation and Selection

We used key evaluation metrics such as Accuracy, Precision, Recall, and F1-Score to determine which was the best model for this project

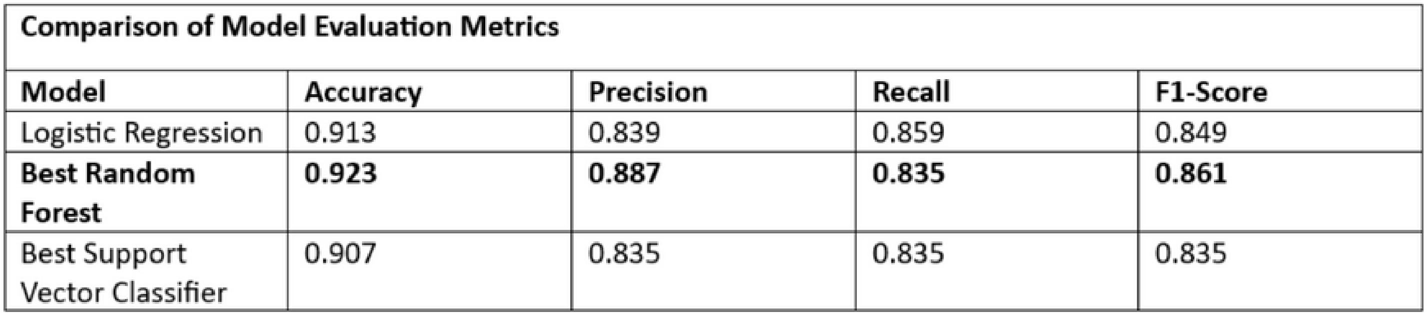
- **Accuracy:** Random Forest achieved the highest accuracy (0.923) which means that it correctly predicts the hiring decisions more often then the other 2 models

- **Precision:** Random Forest also has the highest precision (0.887) which indicates its ability to correctly identify suitable candidates while also minimizing false positives.

- **Recall:** Logistic Regression performs slightly better in recall (0.859) compared to Random Forest (0.835). This suggests that Logistic Regression does a better job with identifying the most suitable candidates, but at the risk of some false positives.

- **F1-Score:** Random Forest has the highest F1-Score (0.861) which balances the trade off between precision and recall. This metrics verifies that Random Forest performs more consistently than the other two models.

Model Evaluation and Selection Cont’d



Comparison of Model Evaluation Metrics

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Logistic Regression	0.913	0.839	0.859	0.849
Best Random Forest	**0.923**	**0.887**	**0.835**	**0.861**
Best Support Vector Classifier	0.907	0.835	0.835	0.835

- The Random Forest model provides the highest Accuracy (0.923) and F1-Score (0.861) which makes it seem as the most balanced choice.
- Ultimately, the best choice is the Random Forest model primarily due to its more balanced performance metrics.
- This project’s focus warrants that we select a model that balances the accuracy of its predictions with minimizing the onboarding of unsuitable candidates.
- The objective of this project requires that a model balances accurate predictions with a focus on minimizing the onboarding of unsuitable candidates. Precision is critical in this case because onboarding unsuitable candidates could result in wasted resources and a drop in overall productivity for a multitude of reasons. The Random Forest model’s strong precision and F1-Score make it the ideal model.
- This model can help organizations confidently predict which candidates are the best fit for the company, which essentially helps to streamline the overall recruiting process.

Recommendations

Leverage the Random Forest Model in Decision Making

- Incorporate the model into the organization's Applicant Tracking System to help recruiters screen candidates based on key features.

Focus Recruitment Strategies on High Impact Features

- Place a larger emphasis on candidates that demonstrate strong interview skill, and personality scores since these features have a strong positive correlation with hiring decisions.
- Steer clear of conservative or moderate recruitment strategies since they demonstrate a negative correlation with successful hiring decisions.

Adjust Candidate Selection Criteria Based on Business Needs

- Leverage EducationLevel data to tailor hiring for specific roles
- Consider the candidate's commuting distance as an effort to incorporate retention planning.

I**mprove Diversity Initiatives**

- Ensure that gender and other demographic features are considered for inclusivity, especially if any imbalance is observed in hiring patterns.

Future Work

Expand the Dataset and Model Features

- Collect more recent data that can take into account shifts in hiring trends (given how much the job market has changed after the pandemic).
- Seek to provide a more holistic evaluation by incorporating features such as TrainingCapabilites, or RemoteWork.

Incorporate Cost Analysis

- Perform a cost analysis breakdown that provides details on the financial impact of having false positives in the hiring process.

Explore more Advanced Models

- Consider implementing ensemble methods such as Gradient Boosting or neural networks.

Collect Model Feedback

- Collect feedback from recruiters and hiring managers to determine how useful the model is to the key stakeholders.

CONCLUSION

IN THIS ANALYSIS, WE IDENTIFIED KEY FACTORS THAT INFLUENCE HIRING DECISIONS. WE BUILT 3 PREDICTIVE MODELS THAT CAN HELP THE RECRUITMENT PROCESS. RANDOM FOREST APPEARED TO BE THE MOST BALANCED, AND THEREFORE THE BEST MODEL TO USE GIVEN THE CONTEXT OF THE PROJECT.

ORGANIZATIONS CAN EXPEDITE THE HIRING/RECRUITING PROCESS BY LEVERAGING THESE MODELS. FUTURE WORK WOULD LIKELY CONSIST OF EXPANDING DATA CAPABILITIES, IMPROVING THE MODEL´S FAIRNESS, AND EXPLORING MORE ADVANCED TECHNIQUES THAT CAN ENSURE THE HIRING PROCESS REMAINS EFFECTIVE.

ULTIMATELY, THE USE OF DATA DRIVEN APPROACHES IN THE RECRUITMENT PROCESS OPTIMIZES DECISION MAKING WHICH CAN POSITIVELY INFLUENCE THE LONG TERM SUCCESS OF AN ORGANIZATION.