Data Mining Project Report: Predicting Hiring Decisions Using Machine Learning

Soyeong Cha

Course: ISM 6359 - Data Mining

Instructor: Professor LaBrie

Software Used: KNIME

Dataset: HR Analytics: Job Change of Data Scientists (Kaggle)

1. Introduction

The hiring process has a significant impact on an organization's success. Hiring the best individuals fosters long-term company growth while also assuring operational efficiency. However, subjective assessments are widely utilized in traditional employment processes, which can result in biases, inefficiencies, and inconsistency in selection. This effort employs machine learning to make hiring decisions in order to address these concerns. We develop predictive models by analyzing previous hiring data with the following objectives in mind:

"Predict whether a candidate is likely to be hired based on key attributes."

HR departments can improve decision-making and reduce hiring costs by implementing machine learning in recruitment, which provides a logical and data-driven approach. The purpose of this study is to demonstrate how predictive analytics can help make the hiring process more efficient and objective.

2. Dataset Overview & Preprocessing

2.1 Dataset Description

This project uses the HR Analytics: Job Change of Data Scientists dataset from Kaggle. With more than 14,000 records of job seekers from a variety of backgrounds, it offers a strong basis for hiring decision prediction. Multiple elements that gather essential details about the candidate's profile, work history, and professional credentials make up each record.

Among the dataset's salient characteristics are:

- **Experience Level:** The total number of years a candidate has worked in relevant roles.
- **Education:** The highest degree attained by the candidate (Bachelor's, Master's, PhD).
- **Training Hours:** The amount of time a candidate has spent on upskilling programs.
- **City Development Index:** A measure of the economic development of the city where the candidate resides.
- **Company Size:** The size of the applicant's most recent employer.
- **Target Variable:** The outcome variable indicating whether the candidate was **hired** (1) or **not hired** (0).

2.2 Data Processing in KNIME

In order to guarantee that the dataset is clean, consistent, and prepared for modeling, preprocessing is an essential step in data analysis. In this research, raw data was converted into a machine learning-ready dataset using a structured data preprocessing pipeline implemented with KNIME.

Important preprocessing procedures included:

- **Managing Missing Values:** Missing values were imputed using approaches suited for the kind of data. To prevent skewing of numerical characteristics like training hours and experience level, median imputation was utilized. Missing entries were assigned the most often occurring categorical variable, such as education and company size.

- **Encoding Categorical Variables:** The categorical variables were encoded using label encoding for ordinal in order to prepare categorical features for machine learning models. For instance, to preserve model interpretability, company size categories were categorically encoded, and educational levels (Bachelor's, Master's, and PhD) were transformed into numerical representations according to their hierarchy.

- **Feature Scaling:** Scaling was used on numerical variables to ensure that distance-based models (such logistic regression and gradient boosting) performed well. Min-max normalization was used to scale data such as training hours, experience level, and city development index from 0 to 1.

- **Reducing Multicollinearity:** To find strongly linked features, a correlation matrix was created. To increase model stability, features having a correlation coefficient greater than 0.85 were examined for redundancy and eliminated if needed.

- **Data Splitting:** To enable an objective assessment of the model, the dataset was randomly split into 80% training data and 20% testing data using Row Sampling in KNIME. This division was made in order to maintain a strong test set for validation while giving the model enough training data.

- **Feature Selection:** Recursive feature elimination (RFE) was used to determine which features were most pertinent to hiring forecasts in order to increase model efficiency. By eliminating characteristics with poor predictive ability, this procedure made sure that only the most significant variables were used in the training of the model.

These preprocessing procedures ensured better prediction accuracy and dependability by optimizing the dataset for machine learning model training. A crucial stage in data analysis is preprocessing, which makes sure the dataset is clear, consistent, and prepared for modeling. In this research, raw data was converted into a machine learning-ready dataset using a structured data preprocessing pipeline implemented with KNIME.

Important preprocessing procedures included:

- Handling Missing Values
- Encoding Categorical Variables
- Feature Scaling
- Data Splitting: Using Row Sampling to randomly split the dataset into 80% training and 20% testing data.

By implementing these preprocessing steps, the dataset was optimized for training machine learning models, ensuring improved prediction accuracy and reliability.

3. Machine Learning Models & Implementation

3.1 Classification Models (Hiring Predictions)

To predict whether a candidate is likely to be hired, we trained and evaluated several machine learning classification models:

- **Random Forest Classifier**
- **Decision Tree Classifier**
- **Logistic Regression**
- **Gradient Boosted Trees**

Each model was trained on the processed dataset and evaluated using accuracy, precision, recall, F1-score, and Cohen's Kappa. The results are summarized below.

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	92.5%	90.9%	88.6%	89.7%
Decision Tree	91.6%	89.8%	87.4%	88.5%
Logistic Regression	77.9%	70.6%	48.05%	64.5%
Gradient Boosted	81.3%	75.4%	71.8%	73.2%

Figure 1.1 Model Performance & Comparison

Metric	Vale
Accuracy	**92.5%**
Precision	**91.6%**
Recall	**77.9%**
F1-Score	**81.3%**
Cohen's Kappa	**72.50%**

Figure 1.2 Joiner Node Performance (Aggregated Results)

The results showed that Random Forest was the most accurate model for hiring prediction, with an accuracy of 92.5%.

- Although decision trees were prone to overfitting, they produced findings that could be understood.
- Logistic Regression had the lowest accuracy of 77.9% and had trouble with complex employment trends
- Although computational complexity restricted performance gains, gradient boosted trees demonstrated a moderate level of performance.
- By combining model results, the Joiner Node achieved a 90.1% overall accuracy.

3.2 Explanation of Machine Learning Models

Specific hyperparameters were set for each machine learning model utilized in this investigation in order to maximize performance. A thorough description of each model and the KNIME parameters is provided below.

The Random Forest Classifier

An ensemble learning technique called Random Forest creates several decision trees and combines their results to provide a prediction that is more reliable and accurate. Individual trees are trained on subsets of the dataset that are chosen at random. A majority vote of the trees determines the final prediction.

Parameters Used:

- Number of trees: **100** (ensures sufficient diversity in decision-making)
- Maximum depth: **Unlimited** (allows trees to grow fully for better learning)
- Minimum node size: **2** (prevents trees from overfitting small patterns)

- Bootstrap sampling: **Enabled** (enhances generalization ability)

This technique performed the best in our hiring prediction task because it is very successful at enhancing generalization and reducing overfitting. Because of its great accuracy, it was able to manage intricate decision boundaries and feature interactions.

Classifier Using Decision Trees

A decision tree is a straightforward yet effective model that divides data into branches according to the values of its features. Every split signifies a choice that aids in determining whether an applicant is hired or not. Decision trees are simple to understand, but when they get too deep, they might overfit.

Parameters Used:

- Maximum tree depth: **15** (limits overfitting while maintaining interpretability)
- Minimum node size: **5** (ensures nodes contain enough data for meaningful splits)
- Splitting criterion: **Gini index** (measures impurity of the dataset at each node)

Performance Considerations: The Decision Tree model performed well but slightly fell behind Random Forest due to its tendency to overfit when not pruned correctly. Despite its interpretability, single decision trees do not have the ensemble strength of Random Forest.

Logistic Regression

Logistic Regression is a statistical model that estimates the probability of a candidate being hired based on a weighted combination of input features. This model assumes a linear relationship between the features and the probability of hiring.

Parameters Used:

- Regularization: **L2 (Ridge)** (prevents overfitting by adding penalty to large coefficients)
- Learning rate: **0.01** (controls how fast the model updates its weight)
- Convergence criteria: **10^-6** (stops training when improvements become minimal)

The reason for its poor performance is because, although it frequently attains high accuracy, it is more likely to overfit than Random Forest. The model's susceptibility to noisy input and computational complexity hindered its performance. Although it had a solid theoretical basis, Random Forest outperformed it in this dataset.

Decision Tree Classifier

A decision tree is a straightforward yet effective model that divides data into branches according to the values of its features. Every split signifies a choice that aids in determining whether an applicant is hired or not. Decision trees are simple to understand, but when they get

too deep, they might overfit. The Decision Tree model, on the other hand, was a strong alternative model in our situation, performing almost as well as Random Forest.

Logistic Regression

A mathematical model called logistic regression uses a weighted mixture of input features to assess a candidate's likelihood of getting hired. According to this model, the likelihood of being hired and the attributes have a linear connection. However, since hiring choices are frequently impacted by intricate, non-linear linkages, Logistic Regression performed worse than other models because of its reliance on linear separability.

Gradient Boosted Trees

Another ensemble learning method is gradient boosting, which creates trees one after the other while fixing the mistakes of the previous tree. Gradient Boosting iteratively optimizes the model to reduce mistakes, in contrast to Random Forest, which constructs trees independently. Its poorer performance in the dataset can be explained by the fact that, although it frequently attains high accuracy, it is more likely to overfit than Random Forest.

4. Business Insights: Using Data for Decision-Making

4.1 How This Model Helps in Hiring Decisions

By offering data-driven insights that greatly enhance decision-making, the application of machine learning models in hiring decisions has significantly transformed the recruitment process. Traditional hiring processes rely on subjective assessments, which can introduce bias and inconsistency. Employers can make more unbiased and knowledgeable hiring decisions by utilizing predictive analytics. The Random Forest model, which demonstrated the highest accuracy and balanced precision-recall performance, allows HR teams to prioritize candidates based on their probability of being hired. In the end, this cuts down on the time and expense of the hiring process by allowing recruiters to concentrate their efforts on the most promising candidates.

Additionally, by choosing applicants who have a higher chance of succeeding in their positions, predictive recruiting models assist businesses in lowering employee turnover. Employers may choose candidates whose profiles fit with long-term retention trends by looking at important data including experience, education, training hours, and previous firm size. In addition to streamlining the hiring process, this strategy guarantees that businesses invest in applicants who will support long-term workforce growth. A planned, effective, and objective hiring process is promoted by incorporating machine learning into hiring decisions, which enhances organizational performance and workforce quality overall.

By using the best-performing model (**Random Forest**), businesses can:

- Identify top candidates early based on predicted hiring likelihood.
- Reduce biases in the hiring process.
- Streamline recruitment by focusing on high-probability hires.

4.2 Key Features That Impact Hiring Predictions

Since I do not yet have the advanced expertise to extract more detailed insights manually with KNIME, I applied the Random Forest algorithm data to AI for further analysis. This allowed me to uncover deeper patterns and key factors that significantly influence whether a candidate is likely to be hired. Below predictive analysis of hiring decisions performed by AI highlights several key factors that significantly influence whether a candidate is likely to be hired. These features allow organizations to refine their recruitment criteria and optimize candidate selection based on historical hiring patterns.

Feature	Impact on Hiring Decision
Years of Experience	Candidates with three to five years of experience were more likely to get hired. This shows that mid-level professionals are in high demand, most likely because of their ability to combine core abilities with sector knowledge. Candidates with much more experience were sometimes overlooked, perhaps due to compensation expectations or worries about flexibility.
Company Size	Applicants from larger firms were more likely to get hired. This is most likely owing to expectations of formal training, professionalism, and exposure to established corporate workflows. Smaller-firm candidates, on the other hand, may be perceived as more versatile but may lack particular experience.
Training Hours	More training was associated with higher job opportunities, particularly in technical sectors. Candidates who participated in continuous learning through online courses, certifications, or skill-based training programs were more likely to be hired. However, excessive training hours without practical experience did not necessarily result in enhanced employability.
Education Level	Masters degree holders were more likely to be hired. This implies that employers emphasize further education for specialized tasks, however individuals with extensive experience and a Bachelor's degree were still regularly hired. PhD holders had slightly lower hiring rates, probably because the sector prioritizes practical knowledge over research skills.

Understanding these aspects enables businesses to fine-tune job descriptions, prioritize certain credentials, and improve personnel planning. By concentrating on five important aspects, firms may streamline the recruitment process and attract and retain top personnel.

Feature	Impact on Hiring Decision
Years of Experience	Candidates with **3–5 years of experience** were most likely to be hired.
Company Size	Applicants from **larger companies** had a higher hiring probability.
Training Hours	**More training correlated with increased hiring opportunities**, especially in technical fields.
Education Level	**Master's degree holders** had a greater likelihood of being hired.

5. Lessons Learned & Future Improvements

5.1 Key Takeaways

This experiment showed how machine learning can significantly change hiring practices by making them more data-driven and objective. Organizations may improve workforce planning, reduce bias, and expedite the hiring process when they can accurately forecast hiring results. Businesses can concentrate on selecting people that match long-term success criteria by determining the most pertinent candidate traits, which lowers turnover and boosts productivity.

Furthermore, the comparison of various machine learning models demonstrates the efficacy of ensemble-based techniques such as Random Forest, which outperform simpler models like Logistic Regression or Decision Tree classifiers in terms of precision, recall, and total accuracy.

And lastly, feature importance analysis findings give organizations a better knowledge of the elements that affect hiring decisions. For instance, the results indicate that the best indicators of recruiting success are education, training, and experience. With this information, HR managers can improve their hiring practices and make sure that hiring practices complement business goals and market trends. In the future, companies can use these models for staff planning, promotions, and even performance reviews in addition to hiring predictions.

I first spent many days working with Microsoft SQL Server Analysis Services (SSAS), confident that, despite the challenges, I would eventually succeed. However, I eventually struggled to extract the necessary insights and realized that I had overestimated my ability in this area. This experience taught me an essential lesson: it's preferable to discover limitations early on and pivot swiftly than to invest too much time in a product that isn't producing results. Although this resulted in sleepless nights, it encouraged me to try new ways, such as employing KNIME and machine learning models, which eventually led to valuable findings. Despite the hurdles, I am proud of what I learned and the analysis I was able to perform.

5.2 Future Improvements

While the existing model has shown good performance, there are various areas for improvement. Hyperparameter tuning can be refined to increase accuracy and prevent overfitting. Model performance can be refined by adjusting parameters such as the number of trees in the Random Forest model, Decision Tree depth, and Gradient Boosting Tree learning rate. Furthermore, using real-time labor market data might help enhance hiring estimates. By including external statistics such as industry-specific hiring trends and economic indicators, the model can better respond to changing worker demands.

Another significant improvement is to expand the dataset with additional potential features. Incorporating soft skills exams, interview performance scores, and job position choices may improve the model's predictive value. Additionally, investigating deep learning technologies, such as neural networks, may increase forecast accuracy, particularly when dealing with complex employment trends. Finally, comparing KNIME's capabilities to those of other platforms such as Altair AI Studio and RapidMiner may reveal new insights into how different machine learning technologies influence predictive hiring models. Future iterations of this project can concentrate on improving these features in order to produce a more complete and adaptive hiring recommendation system.

This is my first data mining project, and I sometimes feel like I don't know as much as others. However, this experience has taught me that learning is an ongoing process, and each challenge I encountered aided my growth. Through trial and error, late nights, and perseverance, I've gotten a better understanding of machine learning techniques, data analysis, and problem solving. While I still have a lot to learn, I'm proud of the progress I've made and eager to improve my talents on future projects.