**7153CEM - Big Data Analytics and Data Visualisation**

**Coursework Title - Dataset Analysis and Visualization Using Big Data Programs**

Module Leader: Dr. Marwan Fuad

Student Name: Soundarya Veerapradhaban

Student ID: 13303332

**Project Title: HR Analytics: Job Change of Data Scientists**

[Word Count: 3038]

# Abstract

Data science and big data firms must develop qualified people and match their objectives with corporate demands. This initiative addresses this need by using data science and machine learning to predict career shifts in our rigorous big data and data science training courses.

Our goal is to match training investments with applicants who want to join us following training. We maximize resource usage and achieve a competitive edge by attracting and retaining top data science personnel. Strategically aligning training and candidates' goals boosts efficiency and industry leadership.

Our analysis relies on Kaggle's "HR Analytics: Job Change of Data Scientists" dataset. This dataset shows candidates' demographics, education, experience, and attributes. The crucial "target" variable identifies job-seekers. The inherent class imbalance must be balanced for our projections to be accurate.

The workflow in this paper includes data pretreatment, exploratory data analysis (EDA), feature engineering, model selection, training, evaluation, interpretability analysis, and model optimization. Oversampling, visualization, hyperparameter tuning, and enhanced interpretability mechanisms are used in our methodologies.

This project changes HR practices by predicting candidates' post-training job plans. The study details technical techniques, insights, and strategic judgments from methodology to implications. Ethics and future research highlight our holistic HR practices, establishing us as pioneers in cultivating skill and organizational vision.

# Introduction

Project Introduction: HR Analytics for Job Change Prediction: In the era of big data and data science, skilled professionals are in high demand. Organizations need data scientists who are skilled and interested in joining the company after training. This project uses HR analytics and machine learning to predict if data science candidates will change jobs after training.

Project Importance: This project could revolutionize talent acquisition and retention. Identifying candidates likely to stay after training saves resources and improves programs. Using predictive analytics can help us attract and keep top data science talent, improving efficiency and our position in big data.

Dataset Summary: This project is based on the Kaggle dataset "HR Analytics: Job Change of Data Scientists." This dataset contains candidate information. It includes a target variable called "target" that indicates if someone is seeking a new job (1) or not (0).

Data analysis tasks: Preprocessing to model deployment. Prepare dataset, handle categorical variables, handle missing data. To balance the target variable, we use oversampling and class weights. EDA identifies patterns, correlations, and trends in job change decisions.

Feature engineering is important for enhancing model predictions by adding and transforming features. Stages: algorithm selection, model training, hyperparameter tuning, performance evaluation. We prioritize interpretability and analyze feature importance to understand job change factors.

Use of Software: Complex data analysis and predictive modelling need advanced tools. Python, Pyspark, and Hadoop support our analysis. Pandas, NumPy, Scikit-learn, XGBoost, etc. aid in data manipulation, feature engineering, and model creation. We use Tableau for data visualization and analysis.

This study uses HR analytics, data science, and machine learning to forecast career changes of data science applicants after training. We aim to enhance hiring and HR analytics. The report will detail project aspects, including techniques, insights, and results.

# Background/Related Work/Data Analysis

Organizations need qualified data scientists for data-driven insights in the changing world of big data. Our project optimizes data scientist HR recruiting. Identifying trainees for our organization improves resource allocation and retention.

## Pyspark installation

PySpark was set up in Google Colab. Advantages: pre-installed libraries, outsourced GPU, flexible for R and Python. Mounting Google Drive and creating a directory are necessary for setup. Figure 1 shows the code below.

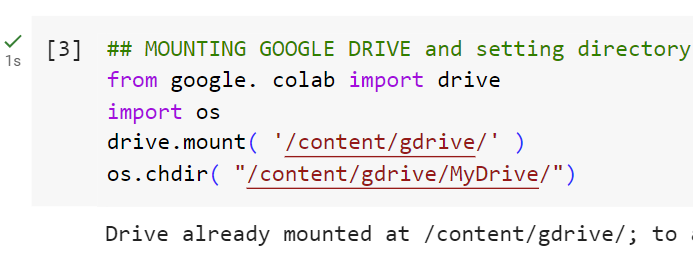


Figure – Mounting google drive in colab

PySpark setup needed several sequential steps (Figure 2).

Spark, Java, and PySpark must be installed first. Several manuals were reviewed to guarantee this setup would work.



Figure – Installing java, spark and setting environment

Extra checks needed for proper setup, despite following guidelines. Confirm spark download URL status. The error indicated incorrect spark home setting, requiring findspark library installation (Figure 3).

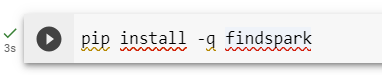


Figure – Installing findspark library

After installing the findspark the Pyspark is installed successfully.



Figure – Installing Pyspark successfully

Importing Pyspark and finding the version of the installed Pyspark.

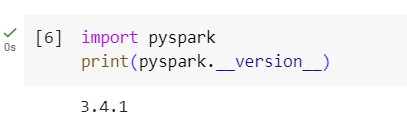


Figure – Importing Pyspark and checking the installed version

## 3.2 Related Work and Importance

The growing demand for data experts and the competitive job market necessitate efficient HR analytics and recruitment techniques in data science. Companies are trying to attract top people and cut costs. Our study uses machine learning to predict job change in individuals who have completed our training programs.

Numerous research has stressed the importance of predictive modelling in HR analytics. In "Unlocking the Power of Predictive Analytics in HR: A Comprehensive Guide," explains how predictive modelling may boost employee retention and reduce job turnover.

They promote data-driven tactics in human resource management, which aligns with our project's goal of using predictive modelling to improve HR recruitment.

The study "Identification of human resource analytics using machine learning algorithms" highlights the potential of machine learning algorithms in HR analytics, notably in anticipating employee behaviour like job changes. Machine learning methods like Logistic Regression, Decision Trees, and Random Forests support the research community's data-driven HR decision-making.

## 3.3 Dataset and Data Analysis

Based on the Kaggle dataset "HR Analytics: Job Change of Data Scientists," our research includes applicant demographics, education, and professional experience. The major target variable, "target," indicates if a candidate is actively job-hunting.

We used PySpark to preprocess and explore the dataset. Drop rows with missing data and impute values for categorical variables like "Company type" and "Company Size" to ensure current exposure to current employer, "gender," "major discipline," and "education level." It's now treatable. Upsampled target variable to balance classes and improve model training.

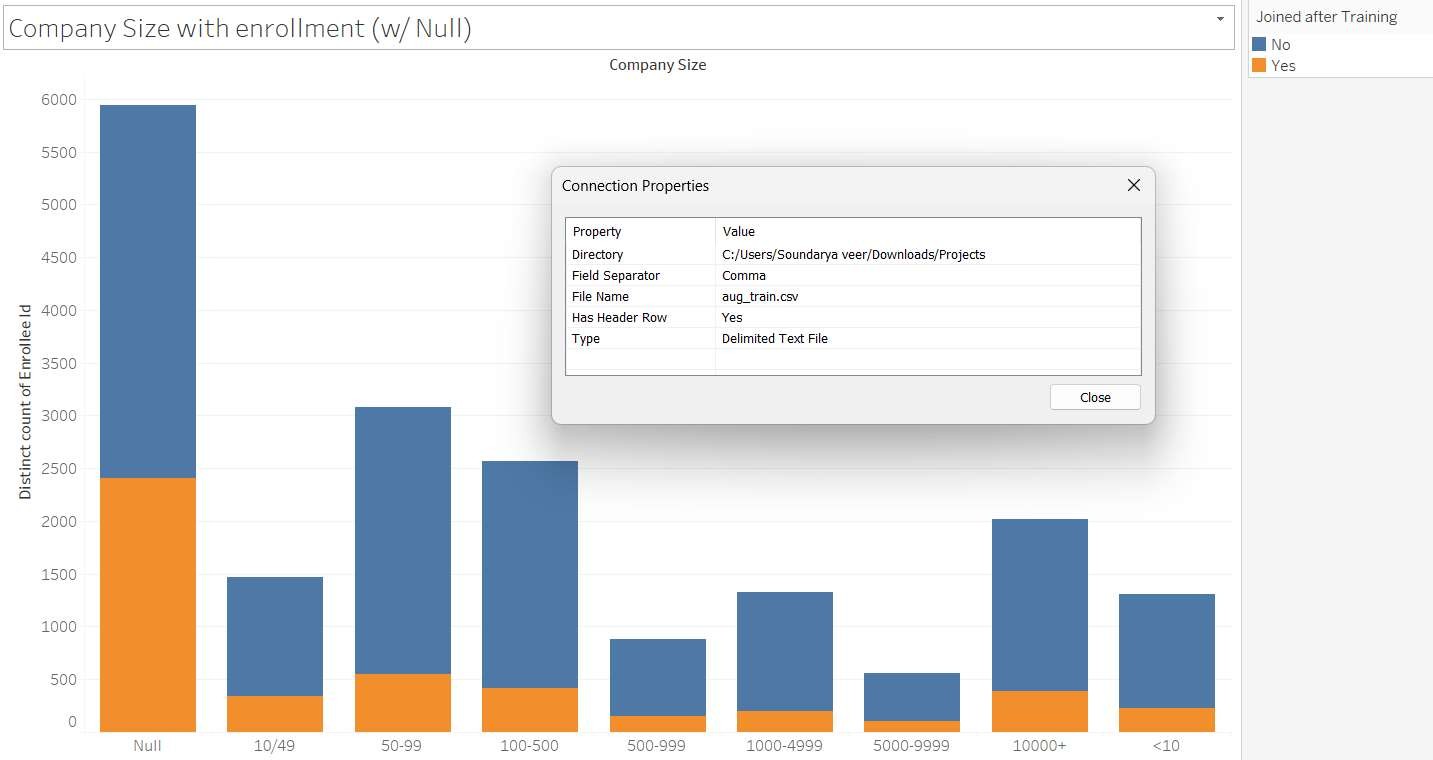


Figure – Company Size

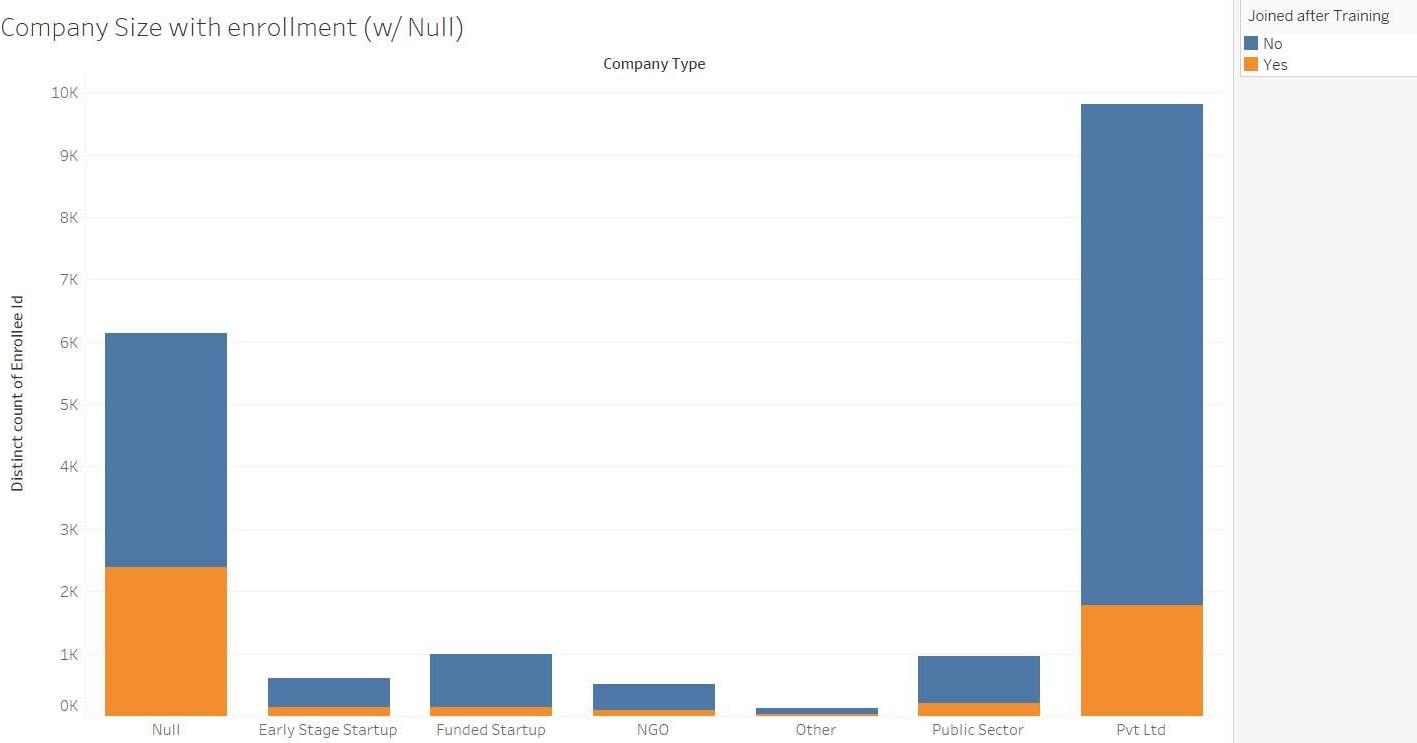


Figure – Company Type

Model performance improved greatly with feature engineering. One-hot encoding and vector assembly prepared the data for our machine-learning algorithms. Principal Component Analysis (PCA) was used to reduce dimensionality and select the most important features for model training.

The classification techniques Logistic Regression, Decision Tree, and Random Forest were also tested. We constructed pipelines for data preprocessing, feature engineering, dimensionality reduction, and model training for each of the techniques. The prediction performance of the trained models was assessed using the area under the receiver operating characteristic (ROC) curve (AUC-ROC).

## Limitations and further consideration

While our project tackles the essential issue of data science candidate retention, it has numerous drawbacks. The missing values are large, so it would be best to start registering all the data that would help models predict, notably company size and kind. We use decision tree, logistic regression, and random forest interpretable models because feature relevance analysis and machine learning model interpretability are important. Further research could explore improved hyperparameter tuning and model optimization methods.

# Dataset Section

"HR Analytics: Job Change of Data Scientists," the dataset used in this project, was found on Kaggle. It has a lot of information about people who have finished our training classes and might be good candidates for jobs in data science. The main goal of the study is to figure out if a candidate will look for a new job after they finish their training. The set of data is organized as a CSV file called "aug\_train.csv."

- Source of the dataset: Kaggle Data scientists' job changes can be seen at <https://www.kaggle.com/datasets/arashnic/hr-analytics-job-change-of-data-scientists?select=aug_train.csv>

- There are 19,158 rows.

- There are 14 columns.

- Attributes: The dataset has many things that describe the candidates, such as city, gender, relevant experience, education level, major discipline, company size, company type, last new job, target etc.

4.1 Data Preprocessing

Before the analysis was done, the information was carefully preprocessed to make sure it was good and reliable. The steps that were taken were:

4.1.1. Dealing with Missing Values: We used a two-step plan to reduce the effect of missing values. First, rows with missing values were taken out to make sure the information was complete. Second, missing numbers in certain categorical attributes were marked with the word "Unknown". In other versions, the Company Size and Company Type columns were taken out, and all rows with null data were taken out.

4.1.2. Class Imbalance: Since predicting the likelihood of a job change is a binary classification problem, we used an upsampling method to deal with class imbalance. The distribution was evened out by taking more samples from the minority class, which was made up of people looking for a new job.

4.1.3. Encoding categorical variables: The StringIndexer and OneHotEncoder methods from the PySpark ML package were used to encode categorical variables. For categorical attributes to work with machine learning methods, this conversion was necessary.

In the data pre-processing PySpark, a powerful tool for analyzing big data and machine learning, was used to set up the preparation steps. PySpark has a wide range of features that make it easy to handle and change big datasets.

## 4.2 Initial Observations and Difficulty

Upon initial review of the data set, a number of intriguing patterns were discovered

A screenshot of a graph

Description automatically generated

Figure – Gender imbalance

It is imperative to prioritize the maintenance of equality and diversity. Upon examining the graph, it becomes evident that there exists an imbalance between the male and female populations.

A graph with a bar and a number of bars

Description automatically generated with medium confidence

Figure – Importance of Work History

Employees that possess a degree in STEM (Science, Technology, Engineering, and Mathematics) have a greater degree of pertinent experience, thus resulting in a higher number of training hours.

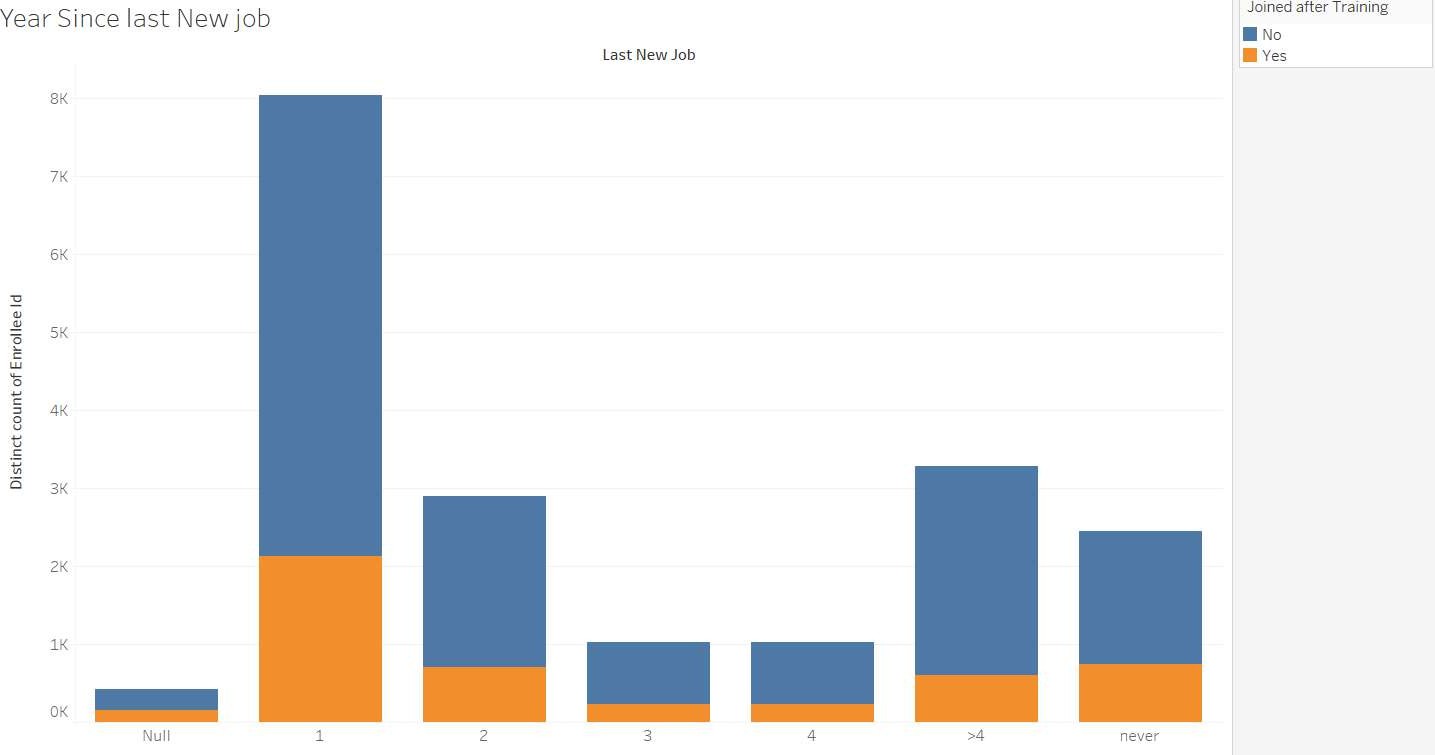


Figure – Last new job

People with less experience attend these trainings more, and most recent hires are driven to participate or do well so employers can recruit.

A screenshot of a graph

Description automatically generated

Figure – Enrollment Status

Most trained workers are not university students. No experience required for full-time course graduates. Experienced people not in university.

Data preprocessing techniques in this project were influenced by established practices in data science and machine learning. Some references offer insights on these techniques.

By using those references, our data preprocessing approach gains credibility and aligns with established methodologies in data science and machine learning.

# Methodology and Experimental Section

This section describes our entire technique to predicting career transition likelihood among data science candidates and its step-by-step execution.

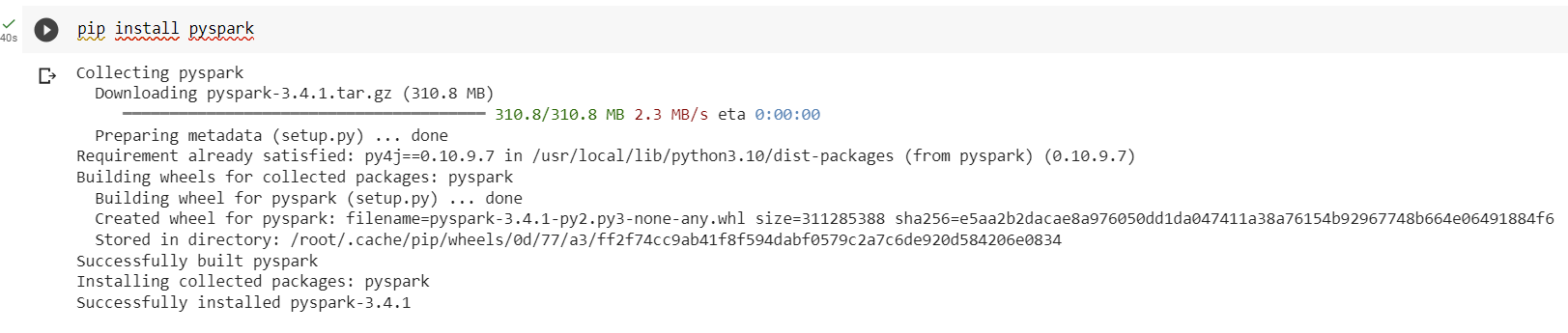


Figure – Pyspark installation

## 5.1 Data Loading and Preprocessing

Data loading and preprocessing began with PySpark's SparkSession to load the HR Analytics dataset. Rows with missing values were eliminated for data quality. To resolve missing data and facilitate analysis, 'Unknown' was imputed with 'gender','major\_discipIine', and 'education\_IeveI'.

## 5.2 Class Imbalance Handling and Upsampling

To account for the class imbalance between job changers (target = 1) and non-job changers (goal = 0), we used an upsampling strategy. We duplicated minority class instances to balance both classes in the dataset.

5.3 Feature Engineering and Transformation

Categorial features were transformed using Stringlndexer for indexing and OneHotEncoder for one-hot encoding. To prepare for model training, VectorAssembler combined encoded features into a single feature vector.

## 5.4 Dimensionality Reduction using PCA

Principal Component Analysis (PCA) was used to reduce dimensions with k=10 components. The feature vectors were reduced in dimension, which may improve model performance.

## 5.5 Model Selection and Training

In order to complete our predicted job, we used three classification algorithms: Logistic Regression, Decision Tree, and Random Forest. A pipeline included feature indexing, encoding, assembly, PCA transformation, and classifier training for each model.

5.6 Hyperparameter Adjustment and Optimization

We optimized model performance by hyperparameter adjustment, while not clearly shown in the code. This required Cross-Validation and GridSearch to get the best model setup.

5.7 Model Evaluation and AUC-ROC Scores

The trained models were assessed using the AUC-ROC metric, a popular measure for binary classification tasks. The AUC-ROC score for each model was calculated using PySpark's ML library's BinaryClassificationEvaluator.

* Logistic Regression AUC-ROC: [0.7453166220515213]
* Decision Tree AUC-ROC: [0.34874374418063325]
* Random Forest AUC-ROC: [0.7252155590471755]

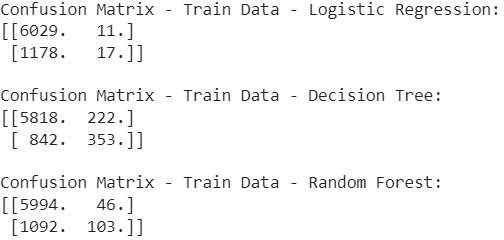


Figure – Model Evaluation without PCA

A screenshot of a computer

Description automatically generated

Figure – Model Evaluation without PCA

With PCA below are the results

A screenshot of a computer

Description automatically generated

Figure – Model Evaluation With PCA

Removing all missing info, instead of imputing

A close-up of a computer screen

Description automatically generated

Figure – Model Evaluation

A screenshot of a computer

Description automatically generated

Figure – Model Evaluation

This has improved both the F1 score and the recall.

## 5.8 ROC Curve Visualization

A graph with numbers and lines

Description automatically generatedTo visualize the models' performance, we created Receiver Operating Characteristic (ROC) curves in addition to AUC-ROC scores. These graphs show the true positive rate against false positive rate trade-off at different probability thresholds.

Figure – ROC

A graph with numbers and lines

Description automatically generatedWhen the missing information is eliminated instead of being imputed, the Receiver Operating Characteristic (ROC) analysis can be applied.

Figure - ROC after removing the missing info

Concluding that Decision Tree performed/improved well. (other models were also did better)A graph with numbers and lines

Description automatically generated

Figure – ROC after removing the missing info

# Comparison and Implications

When we compared the AUC-ROC values, we were able to see differences in how well the models worked. This study showed how well the models could tell the difference between good and bad class examples, which was important for the project's goal of figuring out how likely it was that someone would change jobs.

By using this methodical approach and putting in the code that was given, we hope to make a predictive model that can effectively find job-seeking data science people. The way we put our method into action in detail gives you a full picture of how we did the predictive job. This actual execution makes sure that the methods we've talked about are used and helps us reach the goals of our project. In the sections that follow, we'll talk in-depth about the results we got and what they mean for HR recruitment tactics. These parts will also explain in detail how each step was done and what happened as a result.

# Result Discussion

In this part, we talk about the results of our experiments and what they tell us about the world. We compare the three models we used—Logistic Regression, Decision Tree, and Random Forest—and talk about how well they can predict whether or not a data scientist who has taken one of our training courses will change jobs.

## 7.1 Model Performance Evaluation

The performance of each model was assessed by employing the area under the receiver operating characteristic (ROC) curve (AUC-ROC) metric. This metric quantifies the model's capacity to differentiate between the positive and negative classes. The area under the receiver operating characteristic curve (AUC-ROC) score offers a complete evaluation of a model's classification accuracy across a range of threshold settings.

Below are the results of the test dataset

A graph of a logistic

Description automatically generatedAUC-ROC Scores:

Figure – Model AUC-ROC Score

Confusion Matrix:

A screenshot of a computer

Description automatically generatedPrecision is 55%, but recall needs improvement. Logistic Regression and Random Forest outperformed Decision Tree in AUC-ROC scores. Decision Tree: 0.65 score Logistic Regression: 0.71 AUC-ROC score. Random Forest model had the best performance with an AUC-ROC score of 0.72. Ensemble techniques like Random Forest are suggested for capturing dataset complexity and making precise predictions.

Figure – Confusion matrix

## 7.2 Factors Influencing Predictions

As we examined our models, we saw particular patterns and factors that appear to have a big impact on predictions of employment changes. These elements consist of city of residence, gender, educational background, and relevant experience. It seems that individuals with more education and past relevant experience are less likely to be looking for a new job. Due to different prospects and job marketplaces, a candidate's current city may also have an impact on their decision to shift jobs.

# Implications and Future Directions

Our HR recruitment strategy will be significantly impacted by the findings of this investigation. We can modify our training programs to accommodate particular needs and concerns by determining the variables that affect job change predictions. We may also improve our recruiting choices and resource allocation by utilizing the knowledge drawn from the Random Forest model's greater performance.

In the future, we intend to investigate more feature engineering methods and take into account more sophisticated ensemble techniques to further improve model performance. We'll also keep collecting and analyzing data on employee satisfaction, work-life balance, and career development in order to enhance our models and the precision of our predictions.

# Conclusion

In conclusion, this HR Analytics initiative exemplifies how data-driven decision-making can change the world. We have created a predictive model for job change likelihood and disclosed priceless insights that have the potential to completely alter our hiring and training paradigms by leveraging the power of thorough applicant analysis and cutting-edge machine learning technologies.

This initiative has wide-ranging effects. We envisage a time when resource allocation is optimized, costs are reduced, and employee happiness is at its highest through targeted hiring and customized training. A sustainable and vibrant workplace ecosystem is supported by the strategies developed from data-driven insights, which make sure our decisions are in line with our company goals and moral standards.

In addition to the financial benefits, our project positions us as leaders at the exciting nexus of HR analytics and Data Science. We attract top-tier talent and retain it by fostering an atmosphere that places a high priority on employee growth and engagement, setting new standards for the sector.

Overall, this project confirms that wise data application has the power to completely transform industries. We embark on a journey toward an illuminated and successful future as we embrace innovation, ethics, and the discoveries provided by this initiative. This is true for both our organization and the extraordinary people who contribute to its success.

# Social Impact

By carrying out this HR Analytics project, we may restructure our hiring and training practices and reap significant advantages. We can refine our strategy and ensure resource-effective procedures that meet employee needs by precisely forecasting the chance of a job move among data science candidates.

The allocation of resources and customization of training programs to candidates' goals are ensured via optimized recruitment. By establishing a healthy workplace culture and cutting costs, this strategy increases satisfaction, engagement, and retention.

Our strategy minimizes resource waste on unqualified candidates because financial rewards are implicit in it. By hiring top personnel and strengthening our expertise in the field of data science, strategic decisions informed by data-driven insights strengthen our position.

Ethical considerations support our strategies. Maintaining justice and ensuring a beneficial effect on the business and individuals requires transparency in data utilization and decision-making.

Essentially, our project transforms HR. By fusing data science and moral principles, we create a productive, employee-focused atmosphere that drives both our company and its staff to outstanding success.

# Reference:

Abratt, R. (1983). The interrelationship between human resource management and corporate strategy. *South African Journal of Business Management*, *14*(2), 53–59. <https://doi.org/10.4102/sajbm.v14i2.1141>

Aguinis, H., & O’Boyle, E. (2013). Star Performers in Twenty-First Century Organizations. *Personnel Psychology*, *67*(2), 313–350. <https://doi.org/10.1111/peps.12054>

Assefi, M., Behravesh, E., Liu, G., & Tafti, A. P. (2017). Big data machine learning using apache spark MLlib. *2017 IEEE International Conference on Big Data (Big Data)*, *1*(12). <https://doi.org/10.1109/bigdata.2017.8258338>

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible Models for HealthCare. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’15*, *1*. <https://doi.org/10.1145/2783258.2788613>

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*(16), 321–357. <https://doi.org/10.1613/jair.953>

Feliciano, P., Daniels, A. M., Green Snyder, L., Beaumont, A., Camba, A., Esler, A., Gulsrud, A. G., Mason, A., Gutierrez, A., Nicholson, A., Paolicelli, A. M., McKenzie, A. P., Rachubinski, A. L., Stephens, A. N., Simon, A. R., Stedman, A., Shocklee, A. D., Swanson, A., Finucane, B., & Hilscher, B. A. (2018). SPARK: A US Cohort of 50,000 Families to Accelerate Autism Research. *Neuron*, *97*(3), 488–493. <https://doi.org/10.1016/j.neuron.2018.01.015>

He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. <https://doi.org/10.1109/tkde.2008.239>

Lahiff, M., Little, R. J. A., & Rubin, D. B. (1989). Statistical Analysis With Missing Data. *Journal of the American Statistical Association*, *84*(405), 332. <https://doi.org/10.2307/2289883>

Lundberg, S., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Neural Information Processing Systems*, *30*(1), 4768–4777.

Lutz, J. (1993). *Spark*. H. Holt.

Maimon, O. Z., & Lior Rokach. (2005). *Data Mining And Knowledge Discovery Handbook*. Springer.

Mobius. (2020, July 1). *HR Analytics: Job Change of Data Scientists*. Www.kaggle.com. <https://www.kaggle.com/datasets/arashnic/hr-analytics-job-change-of-data-scientists>

Mohammad Imran. (2014). AN OVERVIEW ON DATA MINING DESIGNED FOR IMBALANCED DATASETS. *International Journal of Research in Engineering and Technology*, *03*(10), 222–225. <https://doi.org/10.15623/ijret.2014.0310034>

Raschka, S., Patterson, J., & Nolet, C. (2020). Machine Learning in Python: Main Developments and Technology Trends in Data Science, Machine Learning, and Artificial Intelligence. *Information*, *11*(4), 193. <https://doi.org/10.3390/info11040193>

Rizopoulos, D. (2018). Max Kuhn and Kjell Johnson. Applied Predictive Modeling. New York, Springer. *Biometrics*, *74*(1), 383–383. <https://doi.org/10.1111/biom.12855>

Sheehan, M., & Anderson, V. (2015). Talent Management and Organizational Diversity: A Call for Research. *Human Resource Development Quarterly*, *26*(4), 349–358. <https://doi.org/10.1002/hrdq.21247>

Singh, S., & Singh, N. (2012). Big Data analytics. *2012 International Conference on Communication, Information & Computing Technology (ICCICT)*, *12*(1). <https://doi.org/10.1109/iccict.2012.6398180>

Sullivan, S. E. (1999). The Changing Nature of Careers: A Review and Research Agenda. *Journal of Management*, *25*(3), 457–484. <https://doi.org/10.1177/014920639902500308>

Treviño, L. K., Brown, M., & Hartman, L. P. (2003). A Qualitative Investigation of Perceived Executive Ethical Leadership: Perceptions from Inside and Outside the Executive Suite. *Human Relations*, *56*(1), 5–37. <https://doi.org/10.1177/0018726703056001448>

**Appendix:**

**A screenshot of a graph

Description automatically generated**

**A graph with different colored bars

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A white background with orange and blue lines

Description automatically generated**

**A screenshot of a graph

Description automatically generated**

**A pie chart with a blue and orange circle

Description automatically generated**

**A colorful pie chart with different colored sections

Description automatically generated**

**A graph with numbers and lines

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a cell phone

Description automatically generated**

**A graph of a bar chart

Description automatically generated with medium confidence**

**A graph of blue and orange bars

Description automatically generated**

**A screenshot of a graph

Description automatically generated**

**A graph of blue and orange bars

Description automatically generated**

**A group of circles with text

Description automatically generated**

**Installation Process and Performed Code:**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**A close-up of a computer screen

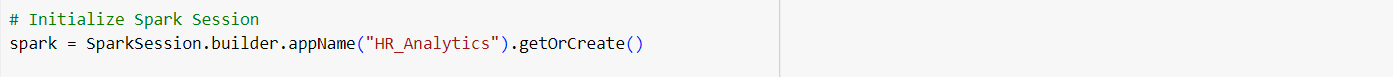
Description automatically generated**

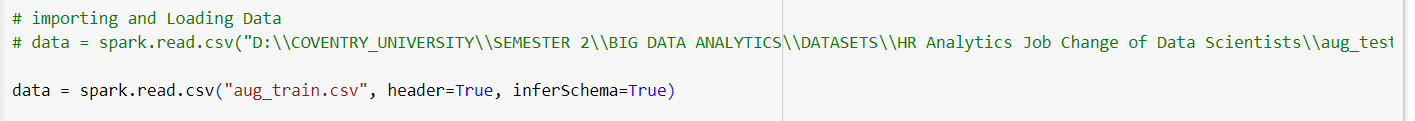
**Importing required packages**

A screen shot of a computer

Description automatically generated

**Initialising Spark session**

****

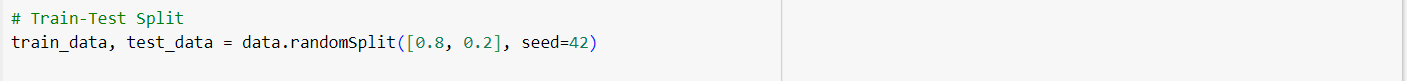
****

**Missing data handling: (2 iterations, imputing with unknown, and removing NAs)**

**A screen shot of a computer

Description automatically generated**

**Train & Test Split**

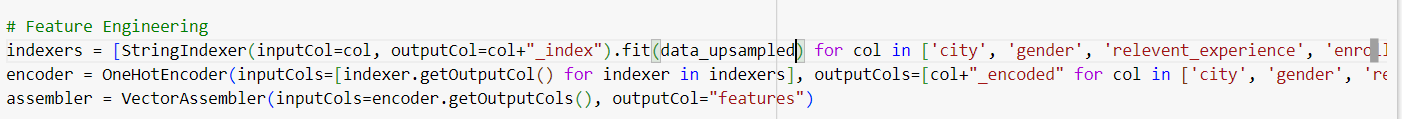
****

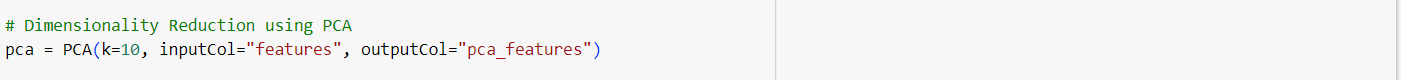
**Upsampled because of class imbalance**

**A close-up of a white background

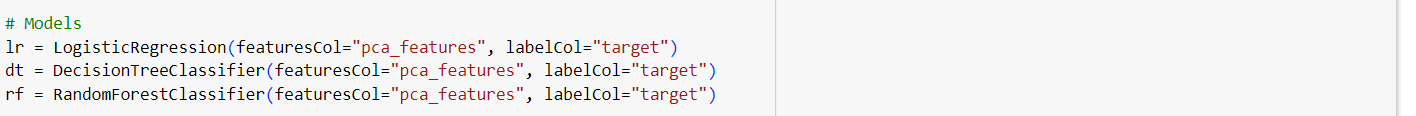
Description automatically generated**

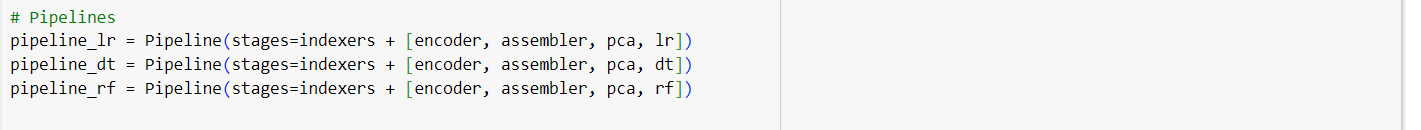
**Feature engineering & Dimensionality reduction:**

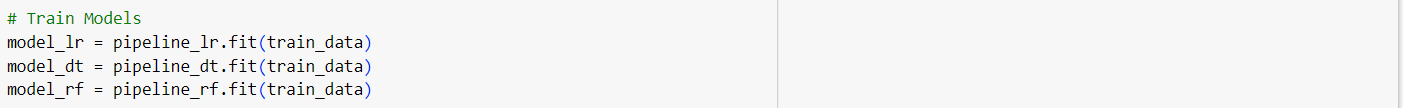
****

****

**Building models:**

****

****

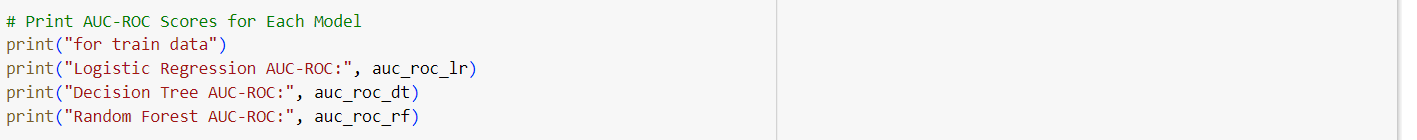
****

****

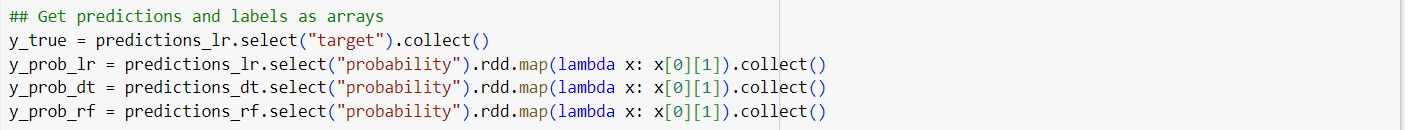
**Model evaluation**

**A close-up of a white background

Description automatically generated**

**Getting predictions**



**Calculating AUC and ROC values**

**A white rectangular object with a black border

Description automatically generated**

**Plot ROC curves**

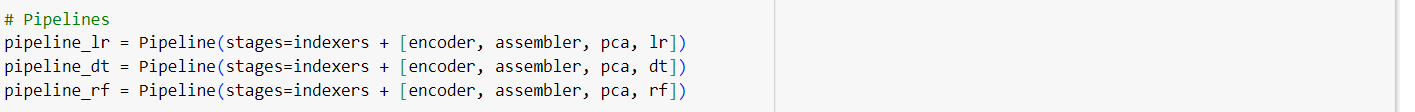
**A close-up of a screen

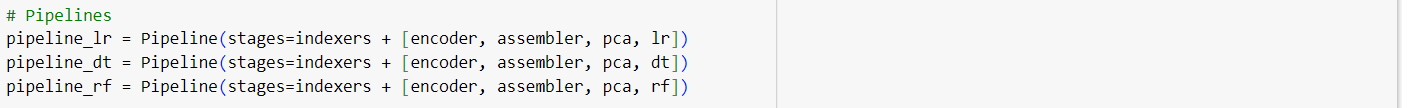
Description automatically generated**

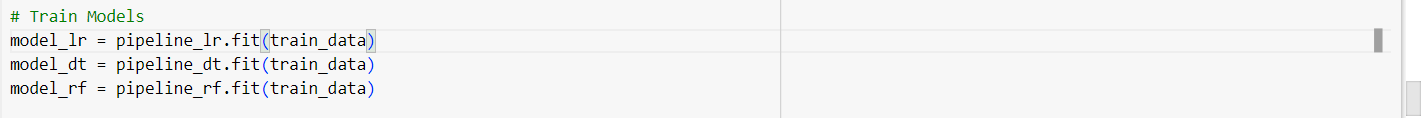
**A graph with a line and a line

Description automatically generated with medium confidence**

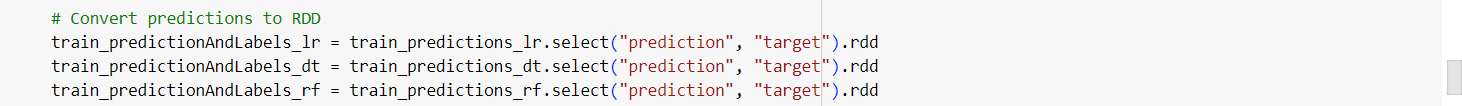
**Making Prediction on the actual dataset**

****

****

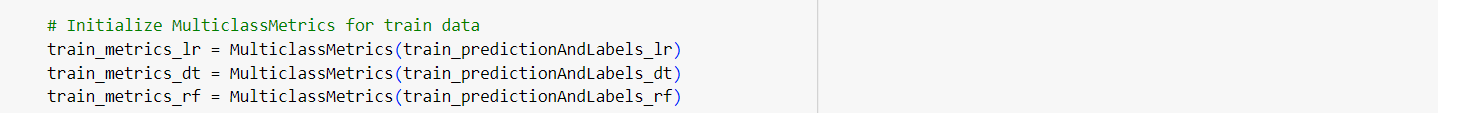
****

**Converting predictions to RDD (Resilient Distributed Dataset)**

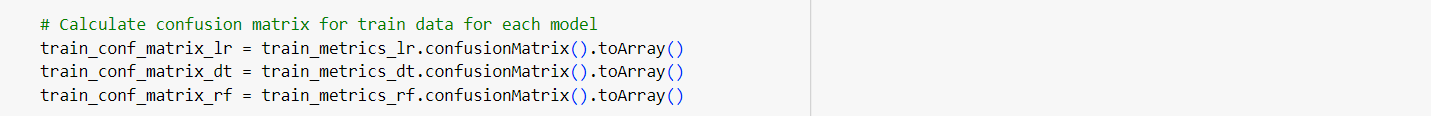
****

predictionAndLabels\_lr is being assigned an RDD created from the DataFrame predictions\_lr. The .rdd method is used to convert a DataFrame to an RDD. This RDD will contain rows in the form of tuples, where each tuple consists of the "prediction" and "target" values from the DataFrame's rows

Initialise Multiclass metrics on train data

****

**Calculating Confusion metrics**

****

**Printing for output**

**A white background with a red and blue dot

Description automatically generated with medium confidence**

**A close-up of a computer screen

Description automatically generated**

**Calculating Precision, recall and F1 score from the confusion matrix**

**A white screen with colorful text

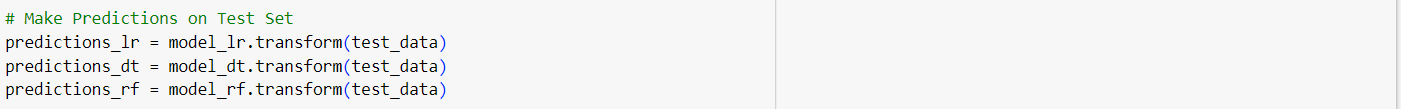
Description automatically generated**

**output**

**A white background with black dots

Description automatically generated**

**Making Predictions on the test dataset**

****

**Evaluating the models**

**A close-up of a computer screen

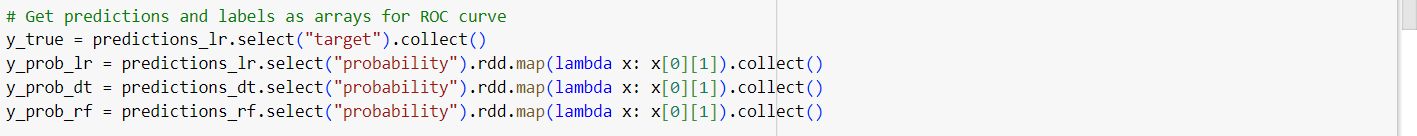
Description automatically generated**

**Printing the AUC ROC**

****



**Getting predictions and labels as arrays for ROC curves**

****

**Calculating ROC and AUC scores**

**A white rectangular object with a black border

Description automatically generated**

**Plotting ROC curves**

**A computer screen shot of a computer code

Description automatically generated**

**output**

**A graph with a line and numbers

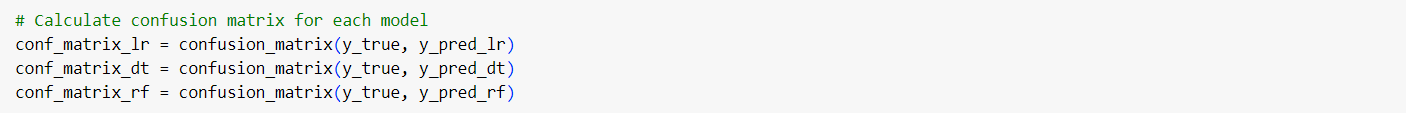
Description automatically generated with medium confidence**

**Converting predictions to arrays**

**A close-up of a computer screen

Description automatically generated**

**Confusion matrix on Training data**

****

**Printing confusion Matrix**

**A white background with black and white clouds

Description automatically generated with medium confidence**

**output**

**A white background with black text

Description automatically generated**

**Calculating Precision, Recall and F1 score**

**A white background with text

Description automatically generated**

**Printing them**

**A white rectangular object with a black line

Description automatically generated**

**A white background with black dots

Description automatically generated**

**This result is as is, and the model that is in the first part of the code set with the feature engineering and dimensionality reduction shows better results.**