Final Course Project
ITM618-021: Business Intelligence & Analytics
Dr. Mehdi Kargar

Marketing Campaign Data Insights:
Portuguese Banking Institution

Khuram Chaudhry

Project Objective

- Provide valuable insights into the effectiveness of different classification models in predicting subscription outcomes based on a dataset.

- Determine whether or not these customers; targeted via a direct marketing campaign, will ultimately open a term deposit.

- Provide insights and analysis for the company on where to target their marketing efforts.

Data Cleaning

- **Deduplication**: Any duplicates identified were removed to eliminate bias and improve performance of predictive models.

- **Outliers:** Identified outliers in key numeric features (age, duration, campaign and nr. employed). Removal of outliers allows for robustness od models and contributes to more reliable predictions.

- **Data Imbalance**: Used a two-step process to achieve a more balanced distribution. This involved converting non-numeric data to numeric formats and handling missing values.

Learning Methods Overview

- Utilized Python libraries extensively for data preprocessing and model training. Key libraries included pandas, numpy, matplotlib, seaborn, and scikit-learn, enabling data handling, analysis, visualization, and machine learning implementation.

- By utilizing a diverse range of algorithms, it ensured that the data was explored from various perspectives to capture potential patterns.

Learning Methods

The primary classification models employed were:

- **Random Forest**: A learning method employing multiple decision trees during training, effective in handling complex datasets, making it suitable for our predictive modeling task.
- **Naive Bayes:** A probabilistic classification algorithm based on Bayes' theorem, particularly efficient and effective for tasks with categorical features. Its simplicity allowed for a pragmatic approach to classification.
- **Decision Tree**: Offer intuitive, interpretable models by recursively partitioning datasets based on attribute values. We utilized decision trees to gain insights into the hierarchical importance of attributes within our dataset.

Learning Methods Overview

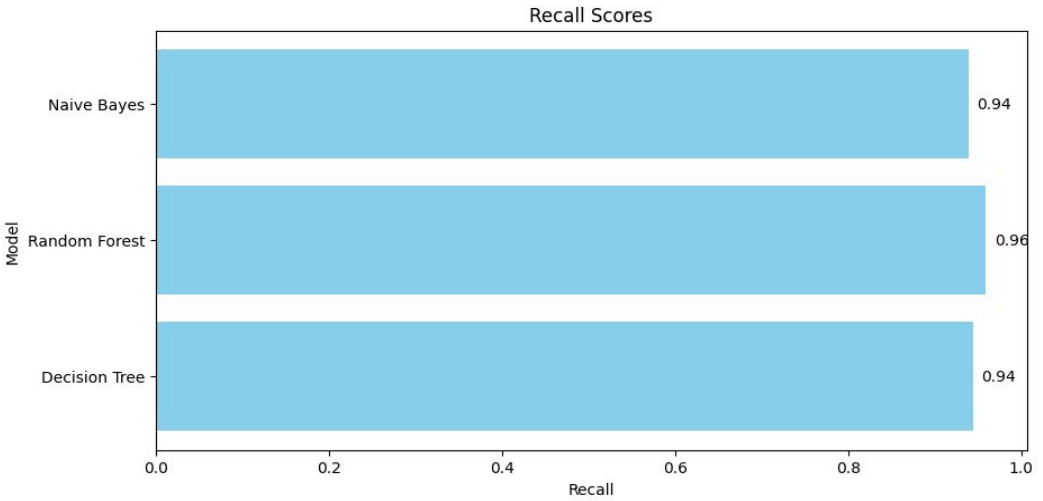
- These learning methods were chosen as they collectively provided a diverse set of tools for building and evaluating predictive models.

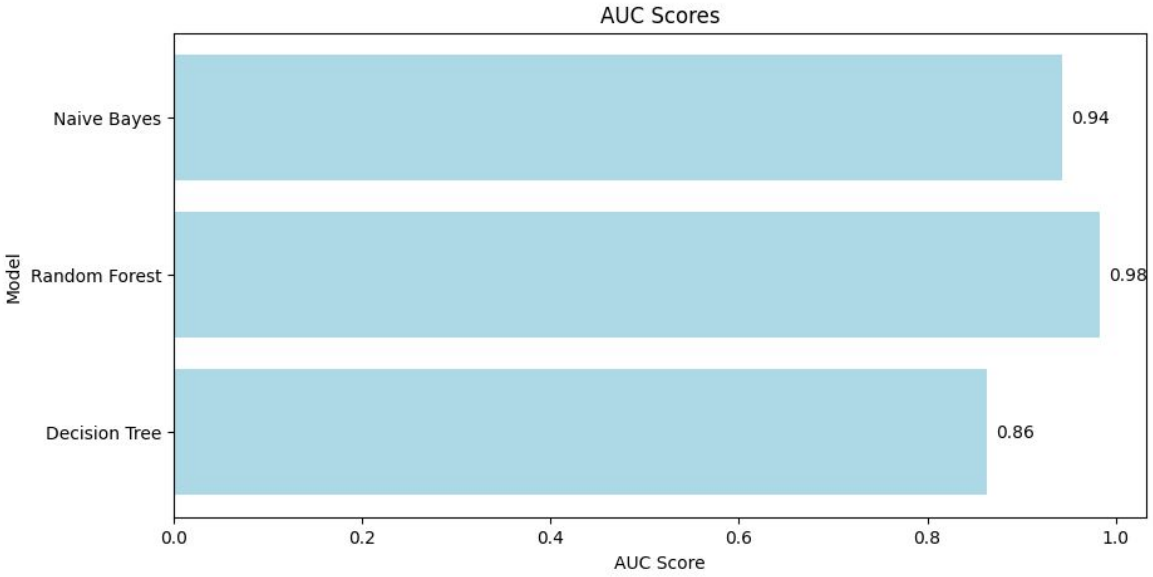
- The rationale for employing multiple models was to ensure a strong analysis, considering that unique algorithms may excel in capturing distinct aspects of the underlying patterns within the data.

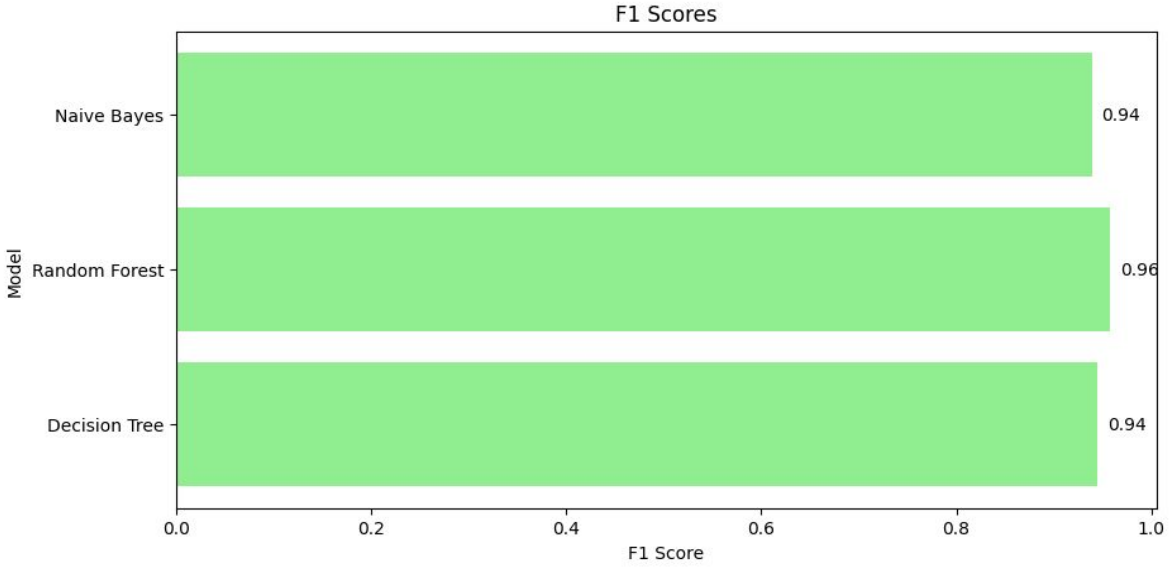
- This approach allowed for the comparison of the performance of different models and selection of the most suitable one for the specific predictive task.

Evaluation

- Random Forest had the highest F1, recall, and AUC.
- Naive Bayes and the decision tree were just a bit lower.
- The decision tree was considerably lower for AUC.

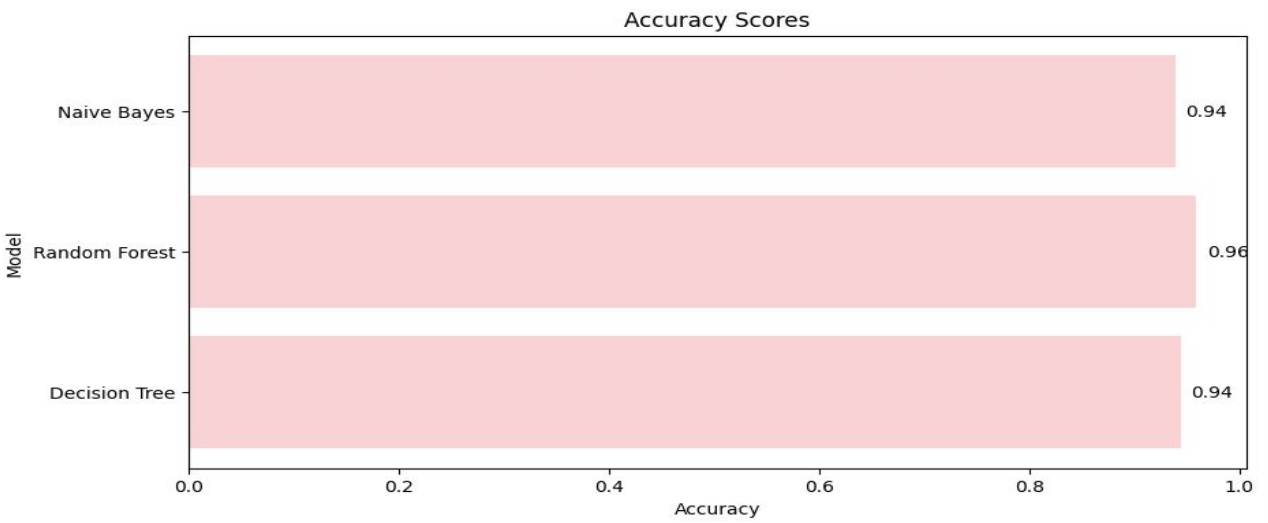


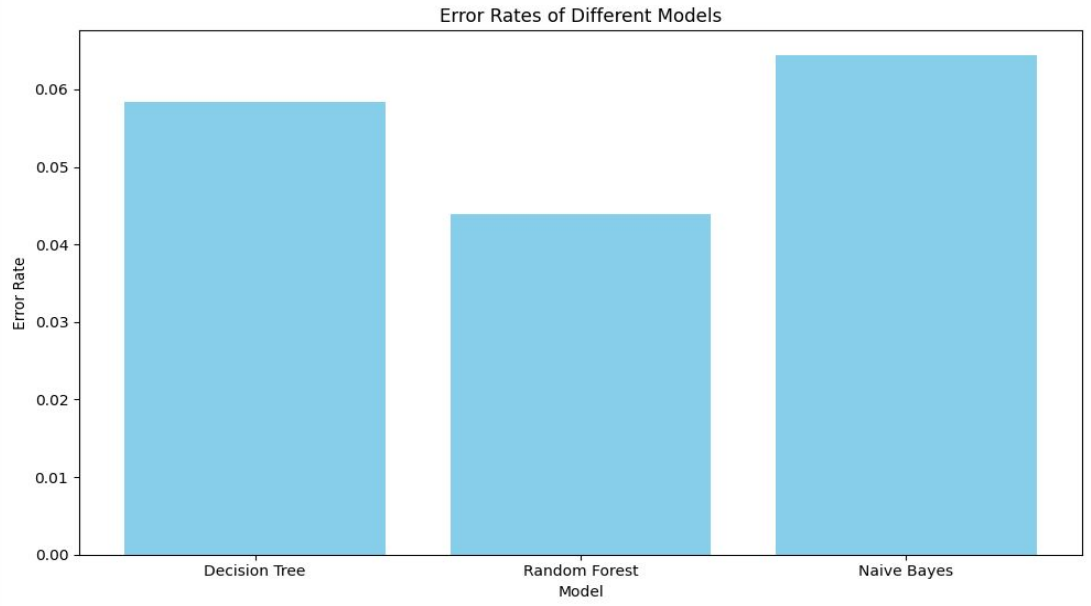


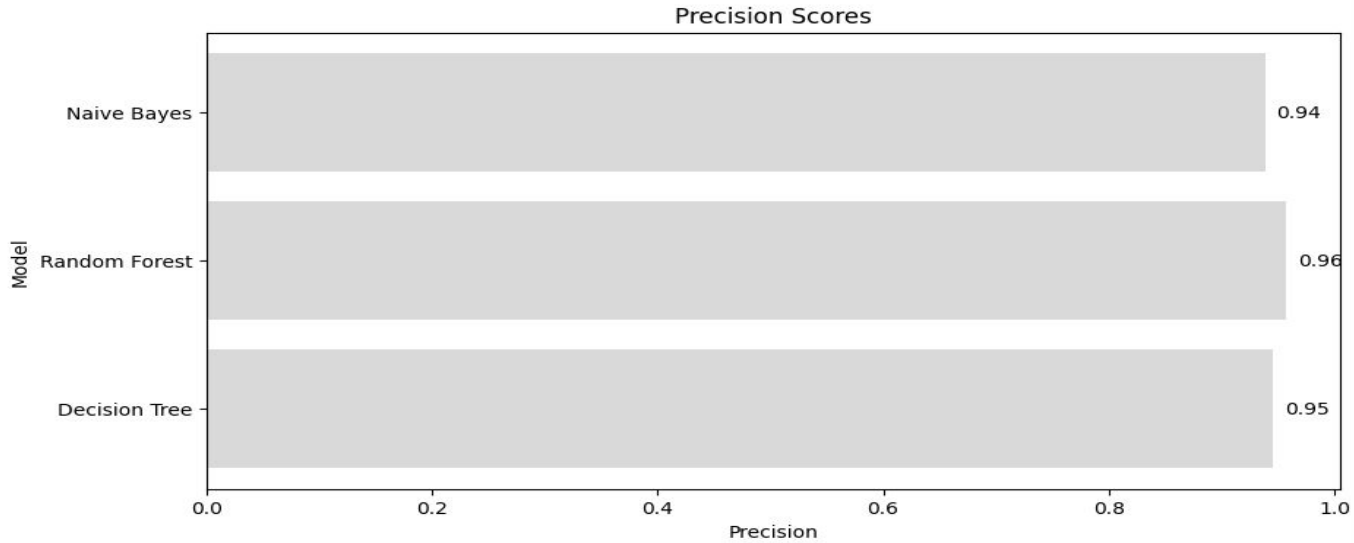


Evaluation

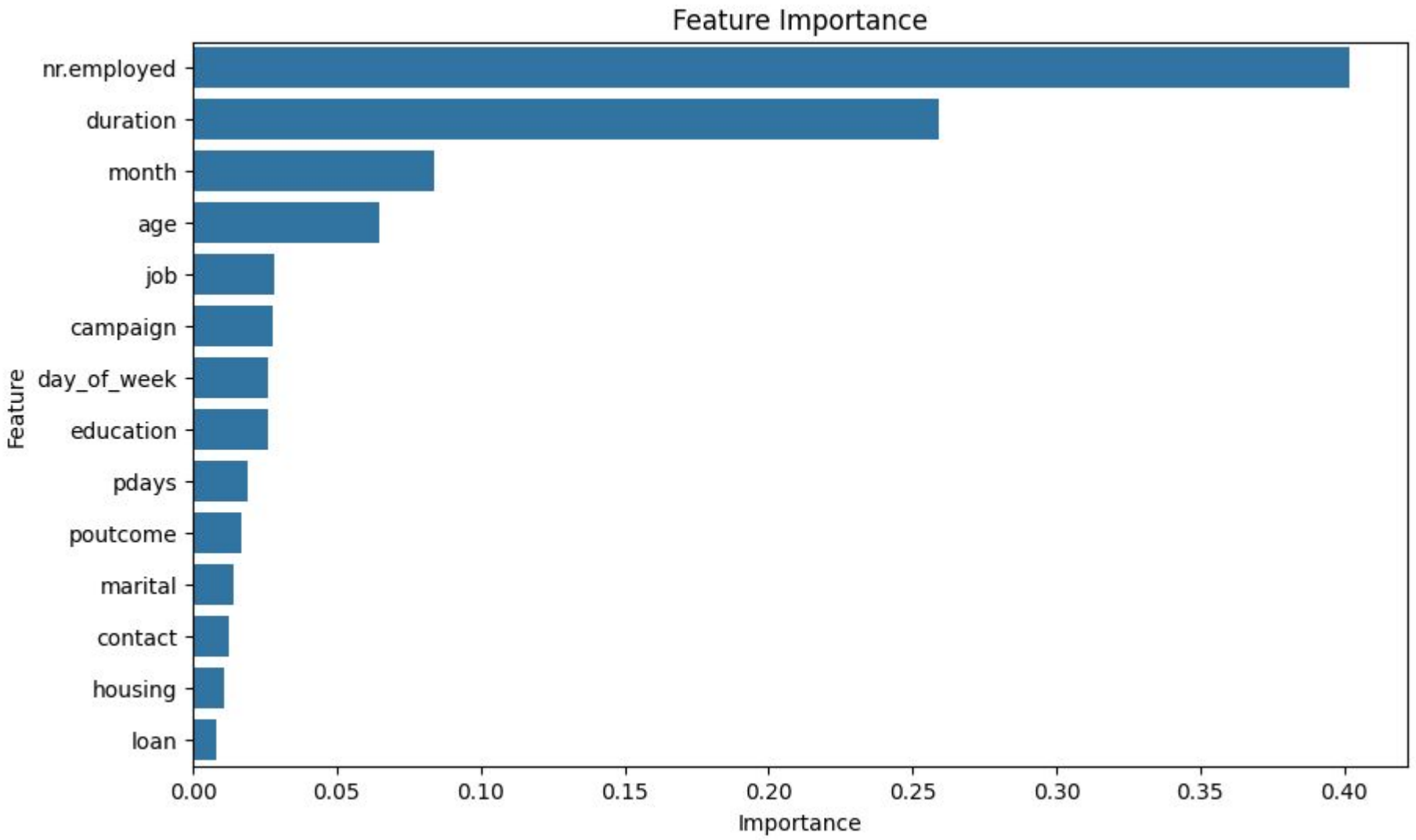
- Random Forest had the lowest error rate and was slightly ahead in precision as well as accuracy.
- The decision tree and Naive Bayes were similar but Naives Bayes had a higher error rate.





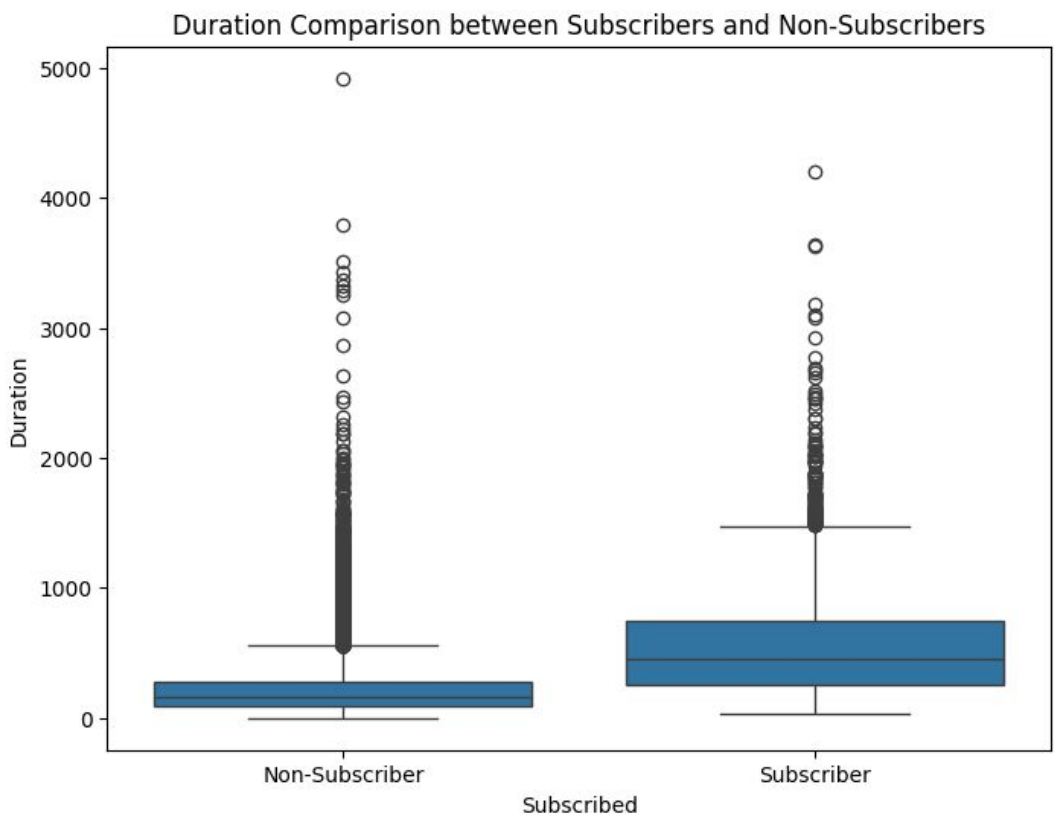


Most Important Features



Contact Duration Insights

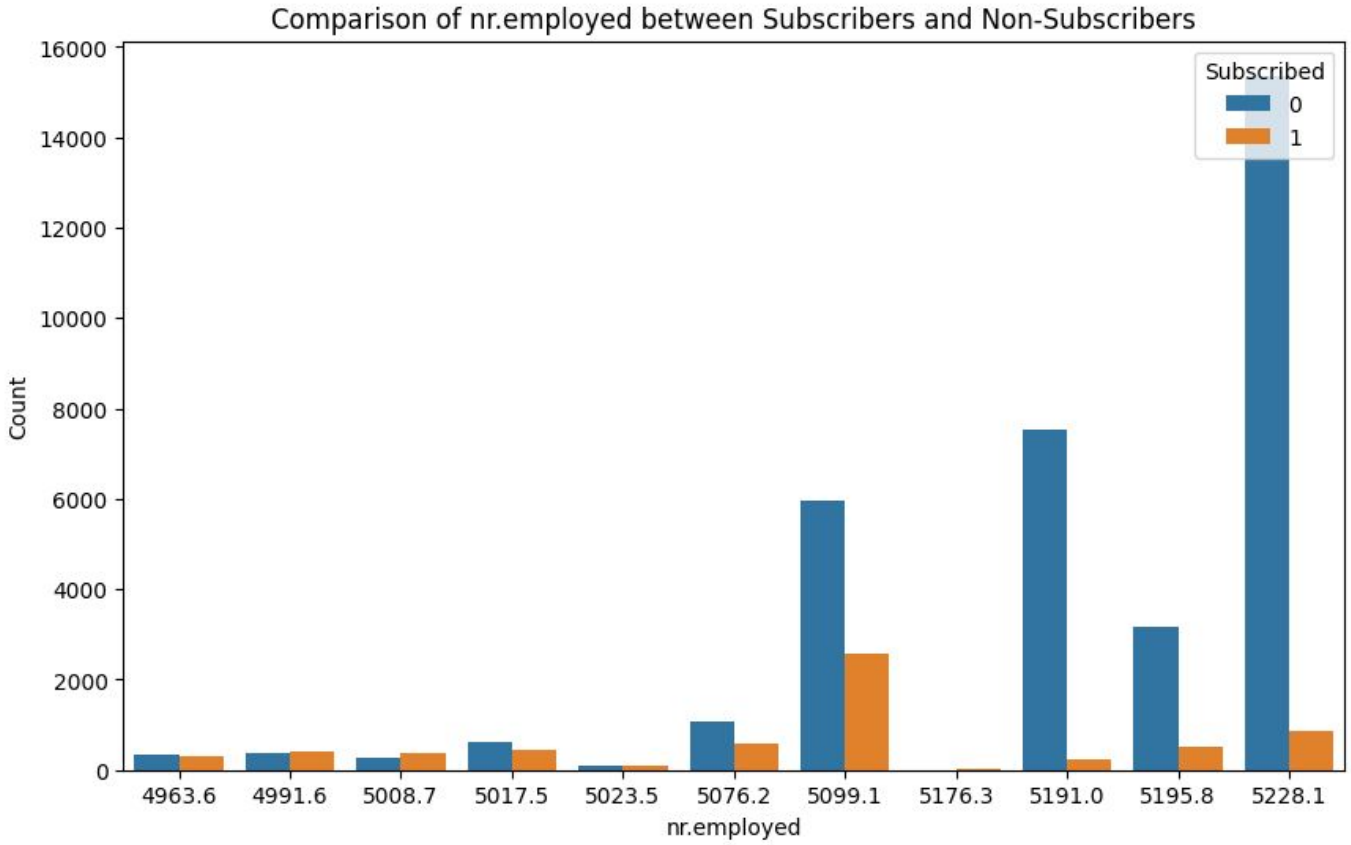
- Longer durations result in higher likelihood of subscribing.
- Investing more time in each customer interaction may increase likelihood of subscription.



Number of Employees Insights

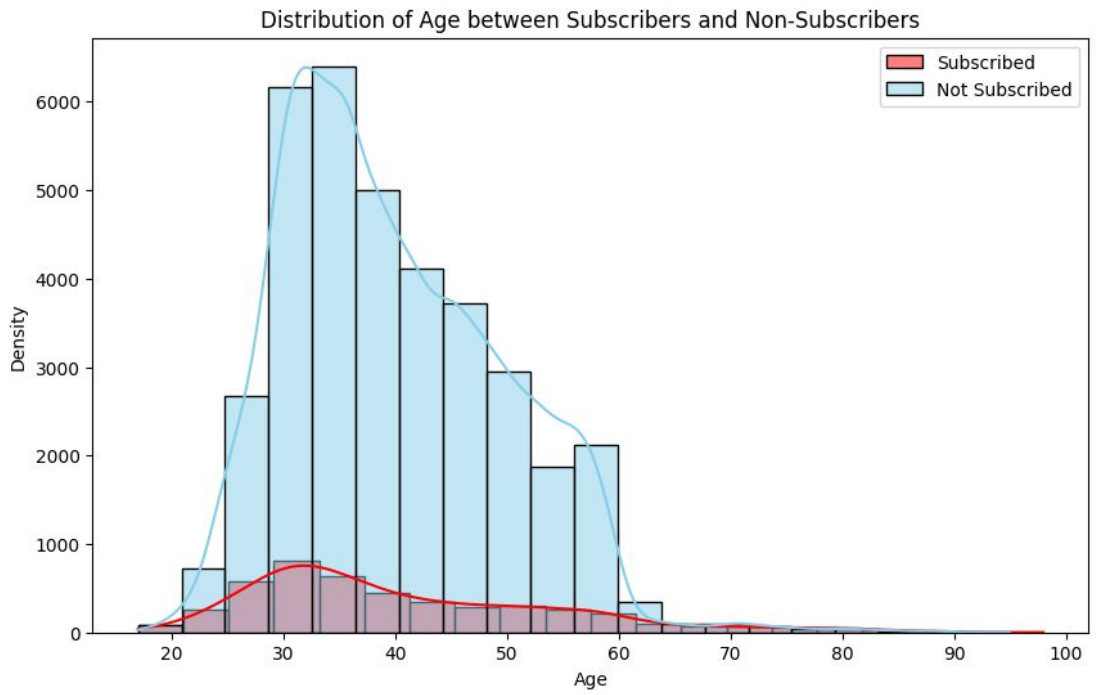
- Number of employees differs significantly between subscribers and non-subscribers. Likely a result of economic conditions.

- Adapting marketing strategies to match the economic conditions of the target audience may boost subscription rates.



Age Insights

- Age levels similar for subscribers and non-subscribers.
- Analyze diverse age groups' preferences and needs to prioritize product development and feature enhancements.
- Customize sales approaches to address specific age-related concerns and preferences, resonating with diverse audiences.



Discussions

- The Random Forest model is the frontrunner, displaying exceptional performance across key metrics - accuracy, precision, recall, error rate, f1 score, and AUC.
- It could potentially be improved by further fine-tuning the hyperparameters and refining feature importance.
- It could also potentially be improved by blending it with other models or using boosting algorithms.
- Duration, number of employees, month, and age are the most important features.