Data Glacier

Your Deep Learning Partner

Exploratory Data Analysis and Predicting Customer Subscription to Term Deposits

Presented By:
Shatabdi Pal
(shatabdi@pdx.edu)
LISUM34

Agenda

Problem Statement

Approach

Dataset Overview

EDA

Business Recommendations

Model Recommendations

Model Performance

Conclusion

Data Glacier

Your Deep Learning Partner

Problem Statement

Problem Statement: ABC Bank aims to predict customer subscriptions to term deposits to optimize marketing efforts and reduce costs. The goal is to identify high-probability subscribers based on past interactions with the bank.

Objective: Develop a machine learning model to predict customer term deposit subscriptions accurately. Optimize marketing efforts by targeting high-probability subscribers and providing actionable insights.

GitHub link: https://bitbucket.org/shatabdi_workpace1/bank-marketing/src/main/

Approach

- Understanding Data and Preprocessing
- Exploratory Data Analysis (EDA)
- Data Cleaning and Transformation
- Handling Imbalanced Data
- Business Insights and Recommendations
- Model Building and Evaluation
- Recommendation based on Model

Dataset Overview

- The dataset consists of categorical and numerical features of a Portuguese banking institution's direct marketing campaigns (phone calls).
- The data is taken from the UCI machine learning data repository.
- The dataset has 21 features, including the label. The target variable y is a binary value ('yes' and 'no')
- It is related to a Portuguese banking institution's direct marketing campaigns.

(phone calls).

Exploratory Data Analysis (EDA)

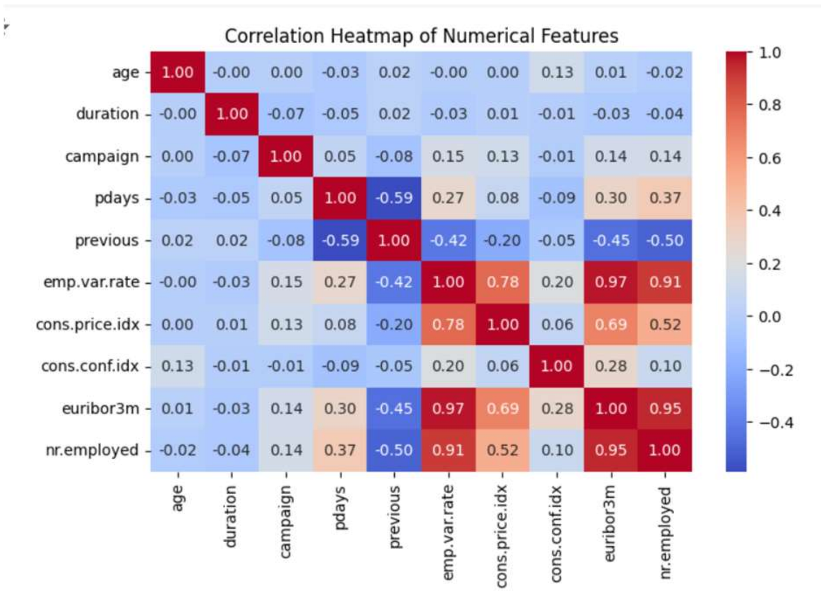
- Exploratory Data Analysis (EDA) involves visually and statistically examining a dataset to extract meaningful insights and understand its underlying patterns and characteristics.

Summary of numerical features

	age	duration	campaign	pdays	previous
count	41176.00000	41176.000000	41176.000000	41176.000000	41176.000000
mean	40.02380	258.315815	2.567879	962.464810	0.173013
std	10.42068	259.305321	2.770318	186.937102	0.494964
min	17.00000	0.000000	1.000000	0.000000	0.000000
25%	32.00000	102.000000	1.000000	999.000000	0.000000
50%	38.00000	180.000000	2.000000	999.000000	0.000000
75%	47.00000	319.000000	3.000000	999.000000	0.000000
max	98.00000	4918.000000	56.000000	999.000000	7.000000

	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
count	41176.000000	41176.000000	41176.000000	41176.000000	41176.000000
mean	0.081922	93.575720	-40.502863	3.621293	5167.034870
std	1.570883	0.578839	4.627860	1.734437	72.251364
min	-3.400000	92.201000	-50.800000	0.634000	4963.600000
25%	-1.800000	93.075000	-42.700000	1.344000	5099.100000
50%	1.100000	93.749000	-41.800000	4.857000	5191.000000
75%	1.400000	93.994000	-36.400000	4.961000	5228.100000
max	1.400000	94.767000	-26.900000	5.045000	5228.100000

Exploratory Data Analysis (EDA)



- The strong positive correlations between emp. var.rate, euribor3m, and nr. employed suggest that these economic indicators are closely related

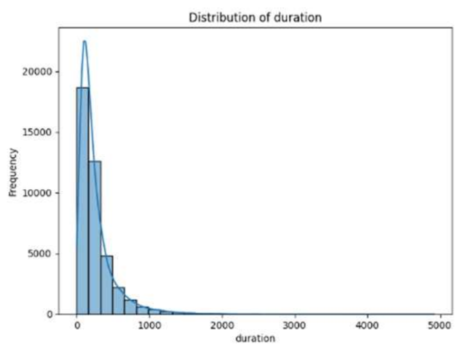
- The negative correlation between emp.var.rate and cons. conf. idx indicates that higher employment variation rates might be associated with lower consumer confidence.

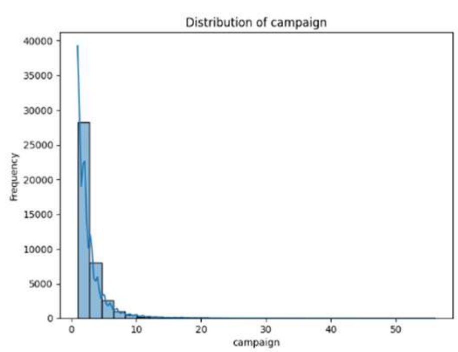
- The relationship between pdays and previous can give insights into how previous contacts impact the time between contacts, which might help understand customer behavior and optimize campaign strategies.

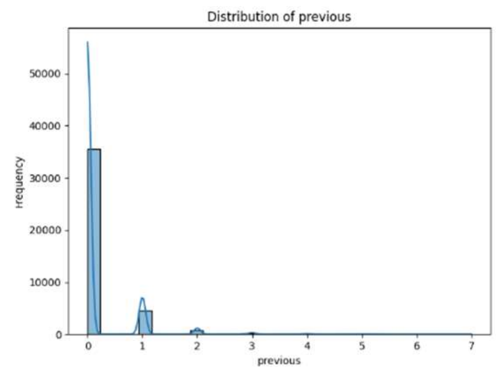
Exploratory Data Analysis (EDA)

- The categorical variables job, marital, education, default, housing, and loan have 'unknown' values, which can be considered missing.

```
Percentage of 'unknown' values in job: 0.80%
Percentage of 'unknown' values in marital: 0.19%
Percentage of 'unknown' values in education: 4.20%
Percentage of 'unknown' values in default: 20.88%
Percentage of 'unknown' values in housing: 2.40%
Percentage of 'unknown' values in loan: 2.40%
```

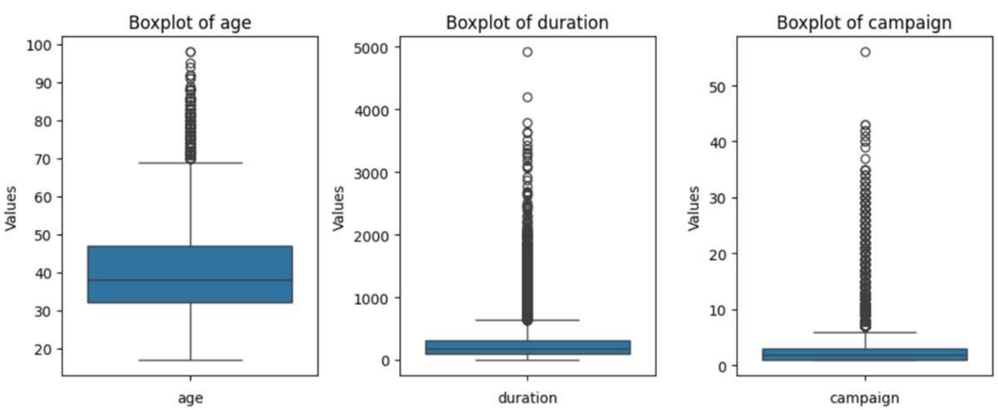




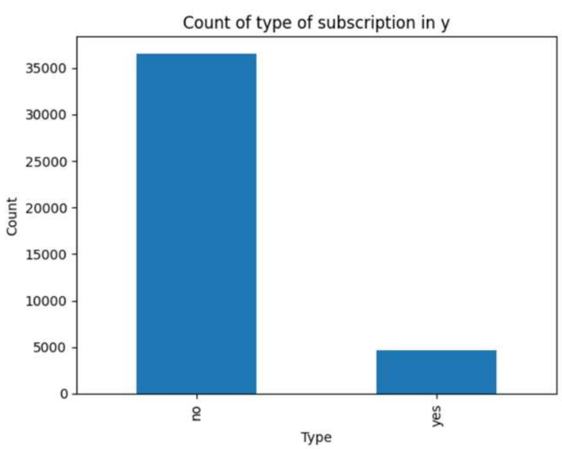


- The skewness of the duration, campaign, and previous features was observed to be more significant than 1.

Exploratory Data Analysis (EDA)



- age, duration, campaign, days, and previous features exhibit outliers that could impact the machine learning model's performance.

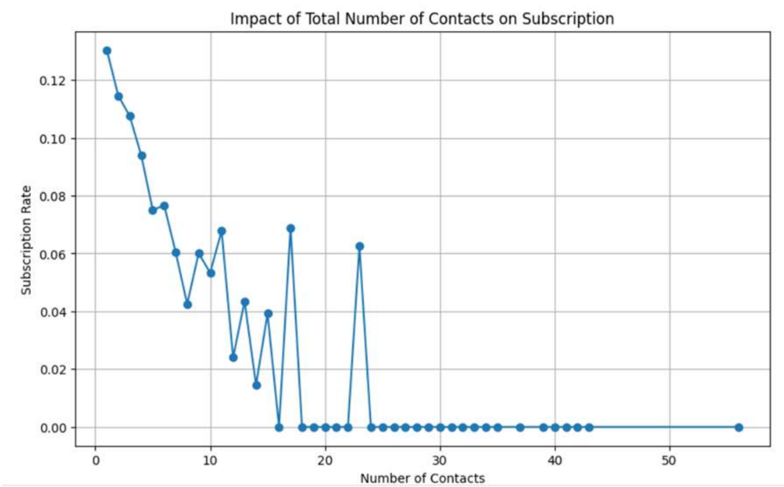


- The target variable was imbalanced, with a higher proportion of 'no' responses than 'yes.’

Data Cleaning and Transformation

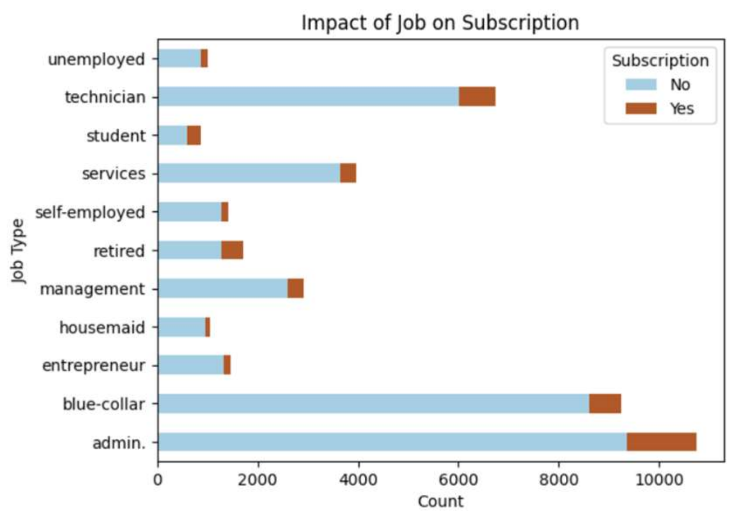
- ‘Unknown’ values were carefully imputed with the most frequent category, ensuring the integrity of the dataset.
- Square root transformation was applied to normalize skewed features.
- IQR (Interquartile Range) dealt with these outliers, ensuring the model's robustness.
- Some new features, such as Total Contacts, Economic Indicators Ratio, and Interaction between features, have been created. These features will be used during model training.

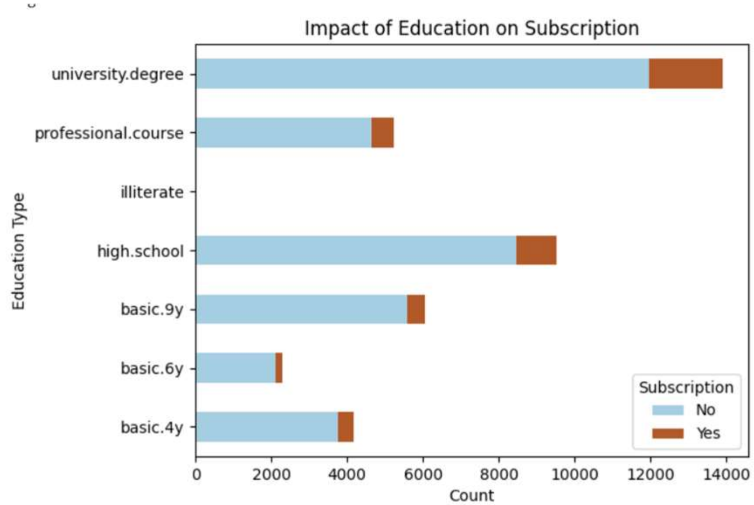
Impact of Total Number Contact on Subscription



- This plot suggests that there might be an optimal number of contacts for maximizing subscription rates.
- The number of contacts increases, the subscription rate tends to decrease.

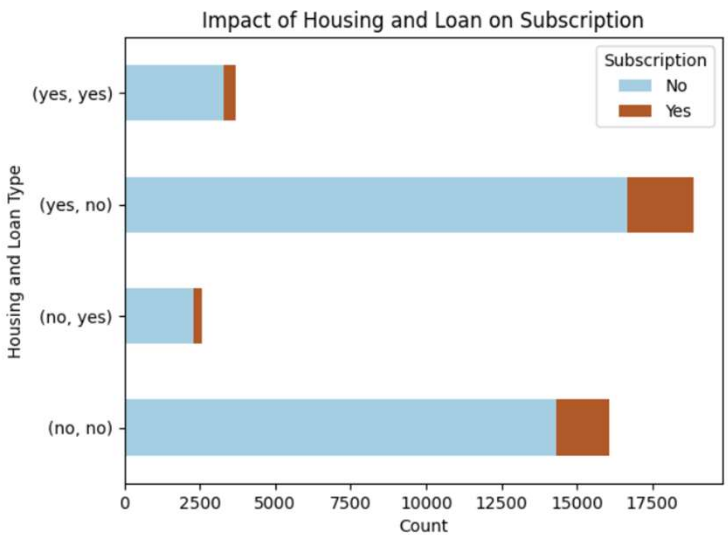
Impact of Job and Education on Subscription





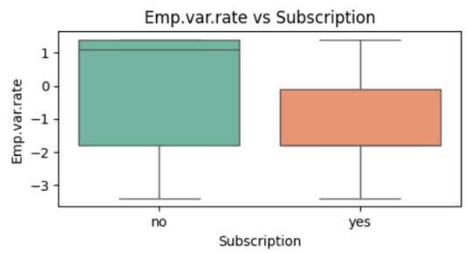
- For most job types shown, the number of people not subscribing is higher than those who subscribe.
- Students and retired individuals have a relatively higher subscription rate than other job types.
- For most other education types, the number of people who do not subscribe is higher than those who do.
- University degree holders dominate in subscribing and not subscribing, underscoring the need for personalized content to cater to their diverse needs.

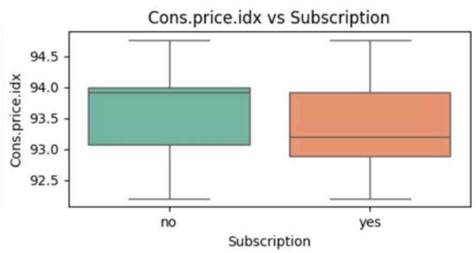
Impact of Housing and Loan on Subscription

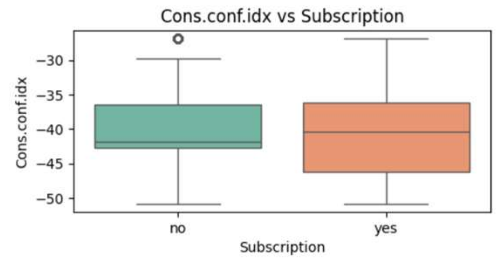


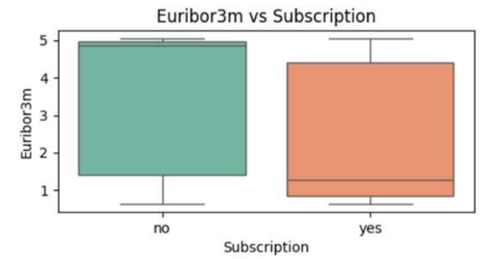
- Having a housing or personal loan may impact subscription behavior.
- Individuals without loans (housing and personal) are more likely to subscribe (as indicated by higher counts).
- Those with both loans (housing and personal) are less likely to subscribe.
- The highest count for ‘No’ subscriptions is in the no housing and no loan category, followed by housing and no loan.

Impact of Social and Economic Attributes





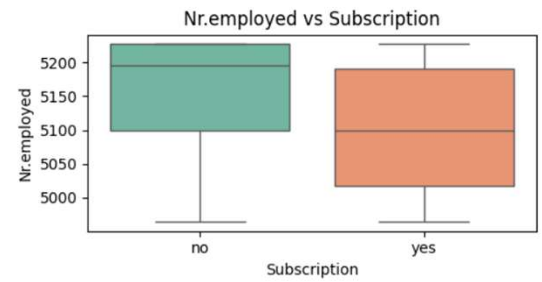




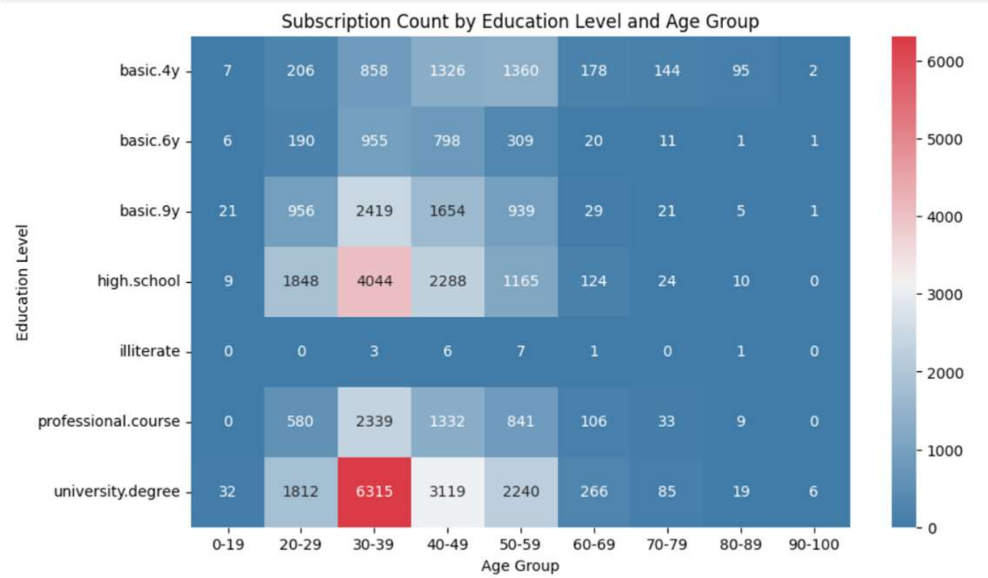
- The employment variation rate (`Emp.var.rate`) and consumer confidence index (`Cons.conf.idx`) show higher medians for non-subscribers and broader ranges, indicating they may influence subscription decisions.

- The consumer price index (`Cons.price.idx`) has minimal impact on subscription decisions, while the Euribor 3-month rate (`Euribor3m`) shows a slight difference in medians, suggesting limited but potential relevance.

- The number of employed individuals (`Nr.employed`) shows similar ranges and medians for both groups, suggesting it may not significantly impact subscription decisions.



Impact of Education and Age on Subscription



- The highest concentration of subscriptions (indicated by darker red shades) occurs in the age group 30-39.

- University degree holders have the highest subscription count, followed by high school graduates and those with professional courses.

Business Recommendation Based on EDA

1. **Optimize Contact Strategy:** Limit the number of contact attempts to prevent customer fatigue and increase subscription rates.

2. **Target-Specific Job Types:** Focus on students and retired individuals due to their relatively higher subscription rates.

3. **Personalize Content for Education Levels:** Tailor marketing messages for university degree holders to cater to their diverse needs.

4. **Address Loan Impacts:** To improve the likelihood of subscribing, offer budget-friendly plans to those with housing and personal loans.

5. **Leverage Economic Indicators:** Time campaigns during favorable economic conditions indicated by Emp.var.rate and Cons.conf.idx.

Business Recommendation Based on EDA

6. Employment Status: Consider other influencing factors, as the number of employed individuals shows minimal impact on subscriptions.

7. Demographic Focus: Focus marketing efforts on the 30-39 age group with the highest subscription concentration.

8. Analyze Non-Subscribers: Investigate reasons behind high non-subscription rates in individuals without loans and address their concerns.

Implementing targeted, personalized strategies based on job type, education level, loan status, and economic indicators, optimizing contact frequency, and focusing on key demographics can significantly enhance subscription rates.

Model Recommendations

- Techniques such as SMOTE (Synthetic Minority Oversampling Technique) can be considered to handle the imbalance of dataset model training.
- Implementation options include logistic regression, ensemble model using random forest classifier, gradient boosting classifier, voting classifier, and Adaboost using decision tree classifier.

Reasoning behind the choice of classifier:

Logistic Regression: This simple and interpretable model is ideal for binary classification. It provides clear insights into feature importance, which helps understand each feature's impact on the prediction.

Random Forest Classifier: This ensemble method reduces overfitting by averaging multiple decision trees. It effectively handles diverse features and provides insights into feature importance.

Model Recommendations

Gradient Boosting Classifier: This classifier sequentially builds models that correct errors from previous models, leading to high predictive accuracy. It is especially useful for handling complex patterns and imbalanced data.

Voting Classifier: This classifier combines predictions from multiple models, leveraging the strengths of different algorithms to achieve higher overall accuracy and robustness in predictions.

AdaBoost using Decision Tree Classifier: Enhances the performance of weak learners by focusing on incorrectly classified instances, making it effective in handling imbalanced datasets and improving prediction accuracy.

Model Performance Metrics

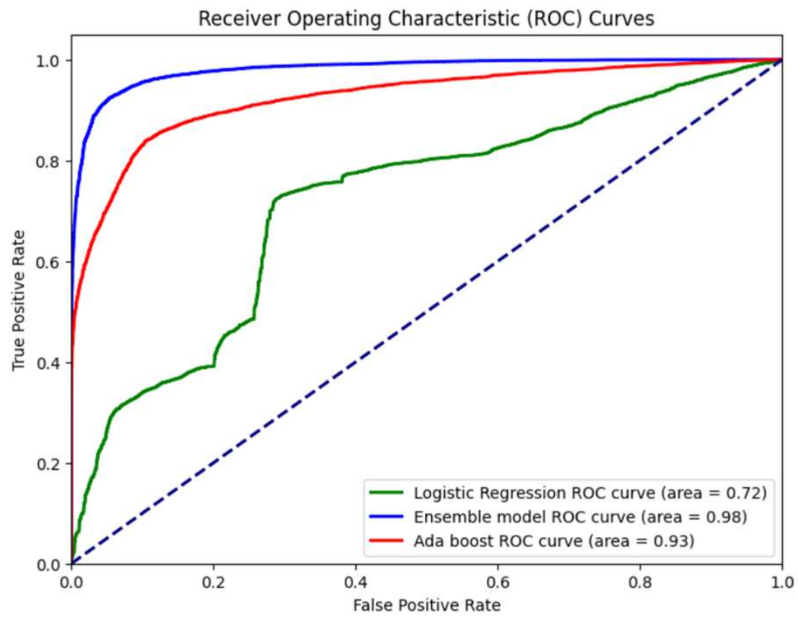
The model's performance is evaluated using a combination of various metrics, such as accuracy, precision, recall, F1 score, and ROC-AUC.

	Model	Accuracy	Precision	Recall	F1 Score
0	Logistic Regression	0.716325	0.713336	0.718243	0.715781
1	Ensemble	0.933266	0.934223	0.931395	0.932807
2	Ada Boost	0.863340	0.896818	0.819499	0.856417

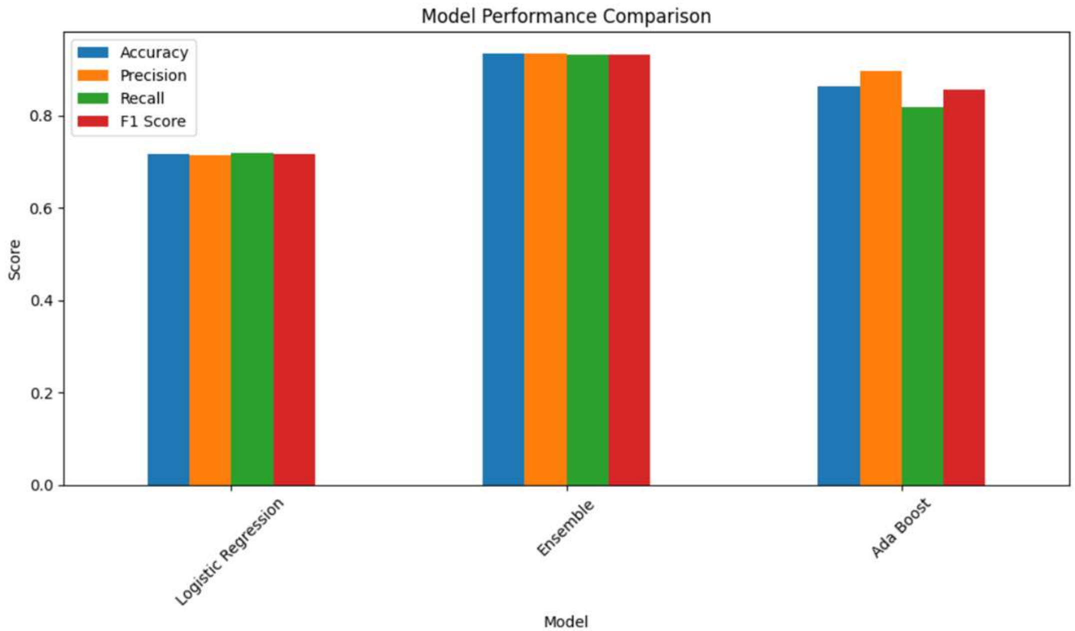
- Accuracy: The ratio of correctly predicted instances to the total cases.
- Precision: The ratio of true positive predictions to the total predicted positives.
- Recall (Sensitivity): The ratio of true positive predictions to the total actual positives.
- F1 Score: The harmonic mean of precision and recall.

Model Performance Metrics

- ROC (Receiver Operating Characteristic) Curve and AUC (Area Under the Curve): The ROC curve plots the true positive rate (recall) against the false positive rate at various threshold settings, and AUC represents the area under this curve.



Model Performance Comparison



The Ensemble model is the most effective for predicting customer subscription, with the highest scores in all performance metrics.

Conclusion

- Target strategies by job type, education level, loan status, and economic indicators to boost subscription rates.
- Focus on students, retirees, and the 30-39 age group, tailoring messages for university graduates.
- Offer budget-friendly plans to those with housing and personal loans.
- Deploy the Ensemble model for best performance; consider Ada Boost an alternative. Logistic Regression is less effective for this application.

Thank You

Data Glacier

Your Deep Learning Partner