Predicting Term Deposit Subscription

John Pole Madhu, Narendhar Babu Yallampalli & Santosh Sai Kadudula

Stetson Hatcher School of Business, Mercer University

BDA 650: Fieldwork

Dr. Wei Xiong

May 3, 2025



Correspondence concerned to this report should be addressed to:
narendharbabu.yallampalli@live.mercer.edu or john.polemadhu@live.mercer.edu

Table of Contents

1. Business Problem/Introduction:

In the banking sector, acquiring new term deposit subscribers is crucial for long-term financial stability and growth. A term deposit is a fixed-term account at a bank where money is deposited for a specific duration, offering higher interest rates compared to regular savings accounts but with limited access to funds until maturity (Chen, 2024). Traditionally, customer outreach efforts rely on random cold calls, leading to inefficient resource utilization, low conversion rates, and potential customer dissatisfaction. These challenges highlight a broader inefficiency in financial product marketing, where banks spend significant resources on campaigns with poor return on investment due to lack of targeted outreach. Furthermore, repeated unsolicited communication contributes to customer fatigue and deteriorates trust, potentially damaging the bank's reputation.

The Portuguese retail bank is not an exception to this and faces a significant business challenge: low subscription rates for term deposits, with only 11.7% conversion from marketing calls. The **core business question** is: How can the bank accurately predict which customers are most likely to subscribe to term deposits, allowing for targeted marketing efforts?"

To address the combined challenges of inefficient marketing and declining customer trust, we propose leveraging predictive analytics to build a data-driven customer targeting model. By analyzing historical call log data from previous marketing campaigns, the model will help identify customers with the highest likelihood of subscribing to a term deposit. Using techniques such as logistic regression and feature importance analysis, we can uncover key demographic, behavioral, and campaign-related factors that influence subscription decisions. This approach enables the bank to shift from broad, untargeted marketing strategies to precise, personalized outreach. Ultimately, it enhances campaign efficiency, improves return on investment, and fosters more respectful customer engagement by focusing only on those most likely to respond positively.

Our objectives for the project are:

Increase Subscription Rates and Marketing efficiency: By predicting and prioritizing customers who are most likely to subscribe, the bank can significantly enhance conversion rates. Success will be measured by comparing subscription rates under the targeted, model-driven approach to the current baseline of 11.7%, highlighting the effectiveness of intelligent outreach.

Enhance Customer Trust and Minimize Communication Overload: By limiting outreach to only those most likely to respond, the bank minimizes unnecessary communication, preserving customer goodwill.

Achieve Growth Targets Sustainably: With improved targeting and resource utilization, the bank can meet its term deposit growth objectives without increasing marketing expenditures.

2. Data Understanding/Data Dictionary:

This dataset contains records from direct marketing campaigns (phone calls) conducted by a Portuguese banking institution. It includes customer demographics, call details, and economic indicators, with the goal of predicting whether a client will subscribe to a term deposit (variable **y**) (Moro, Cortez, & Rita, UCI archives, 2012). The dataset has `16 variables` without the `y` variable and `45,211 instances`. The data has no null values, and the data is consistent across the fields. The detail data dictionary is in the tables.

S.No	Column Name	DataType	Info.	Range/levels
1	Age	Numeric	Client Age	18-95
2	Job	text/categorical	Type of job	12
3	Marital	text/categorical	Marital status	3
4	education	Text/categorical	Level of education	4
5	default	Binary	Has credit in default	1/0
6	balance	Numeric	Average 12month bank balanceEUR	-8019 – 102127
7	housing	Binary	Has a housing loan?	1/0
8	loan	Binary	Has a personal loan?	1/0
9	contact	categorical	Communication type	3
10	day	Categorical	Last contact day of the month	31
11	month	categorical	Last contact month of the year	12
12	duration	Numeric	Last contact duration in seconds	0-4918
13	Campaign	Numeric	Number of contacts performed during **this campaign and** for this client	1-63
14	pdays	numeric	Number of days passed by after the client was last contacted from a previous campaign (-1 means never contacted)	-1 - 871
15	previous	Numeric	Number of contacts made with the client before **this current campaign**, including all prior interactions	0-275
16	poutcome	categorical	Outcome of the previous marketing campaign	4

Table 1: Data Dictionary

In the above data dictionary tables, **range is for numerical data and **levels** is for categorical data.

3. Literature Review/Domain Knowledge

We have reviewed a few peer-reviewed articles that researched this project data. One of the scholarly papers was published by the faculty at the University of Lisbon that addressed this from a socio-economic perspective. The authors used the combination of feature selection and feature engineering. In the feature selection, they used a semi-automatic procedure, i.e. selecting variables based on business knowledge and eliminating unnecessary variables with forward selection. Secondly, they incorporated feature engineering i.e. the addition of socio-economic factors such as `emp.var. rate` (Employment variation rate, with a quarterly frequency 0.929), `cons.price.idx` (Monthly average consumer price index), cons.conf.idx (monthly average consumer confidence index), `euribor 3m` (daily three-month Euribor rate), nr. employed (quarterly average of the total number of employed citizen). Finally, the paper compares the performance of four popular classification models: `logistic regression` (LR), `decision trees` (DTs), neural network (NNs) and `support vector machines` (SVMs). Among these models, `Neural Network` performed the best with AUC = 0.929 (Moro, Cortez, & Rita, A data-driven approach to predict the success of bank telemarketing, 2014).

4. Methodology

This research employs the `CRISP-DM` methodology, a standard framework for data mining projects, commencing with business understanding. The core challenge is the Portuguese retail bank's low 11.7% term deposit subscription rate from untargeted marketing calls. Our primary objective is to develop a predictive model to identify likely subscribers, thereby enhancing marketing efficiency and customer engagement (Domingos, 2012); (Shearer, 2000). This initial phase ensures a clear focus on the business problem and desired outcomes.

The subsequent Data Understanding phase involves a thorough examination of the provided dataset, detailed in the data dictionary. This includes analyzing customer demographics, call specifics, and economic indicators (Han, Kamber, & Pei, 2012). By understanding the characteristics and quality of this data and drawing upon existing knowledge in banking analytics

(Berry & Linoff, 2004), we lay the necessary groundwork for effective data preparation and subsequent modeling efforts aimed at predicting term deposit subscriptions.

5. Data Cleaning

5.1 Null Values

We analyze missing values across 45,211 rows and find that several columns have significant null values. The `poutcome` column has the most missing data (36,959 values, ~81.8%), and `contact` is also heavily affected (13,020 values, ~28.8%). Since the missing data in `poutcome` is too high, we drop the column entirely. In other columns like `'job,' 'education,'` and `'contact,'` we imputed using knn imputation. This cleaning process helps us maintain a structured dataset, minimizing data loss and preparing it for further analysis and modeling.

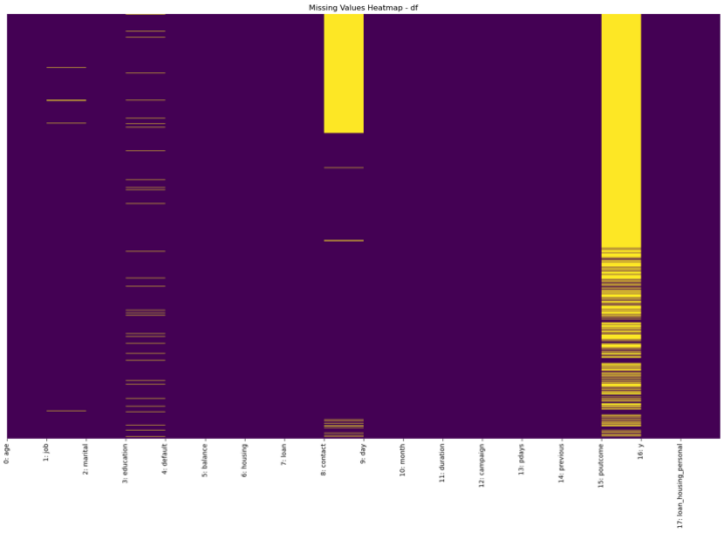


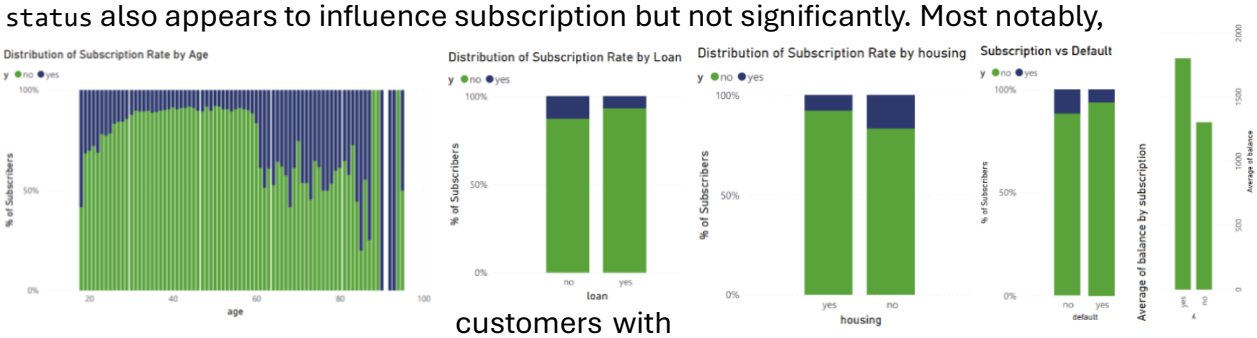
Figure 1 Null Values - 'unknown'

Column	% of Missing Values
1. Job	0.64 %
2. education	4.11 %
3. contact	28.80 %
4. Poutcome	71.75 %

Table 2 % of Missing Values in the dataframe

5.2 Exploratory Data Analysis

The Exploratory Data Analysis (EDA) reveals several insights into the factors influencing the target variable (subscription). We used the Microsoft Power BI tool to visualize the data. Regarding *client data* (age, job, marital status, education, etc.), the subscription rate varies across different demographics. `Age` plays a role, with subscription rates differing across age groups. The type of `job` and the level of education also show variations in subscription rates, indicating that certain professions and education levels are more likely to subscribe. `Marital status` also appears to influence subscription but not significantly. Most notably,



customers with no – `default` status and higher `average balance`, `higher duration`, no housing, and no `loan` are more likely to subscribe.

The likelihood of subscription is influenced by several factors related to the last contact campaign and other client information. The method of contact and month of contact all demonstrate variations in subscription success, with the duration of the last contact being particularly notable, showing a positive correlation between longer durations and higher subscription rates. Additionally, subscription rates are affected by the number of contacts during the current campaign and are strongly predicted by the outcome of the previous marketing campaign, where a 'success' greatly increases the chance of subscription.

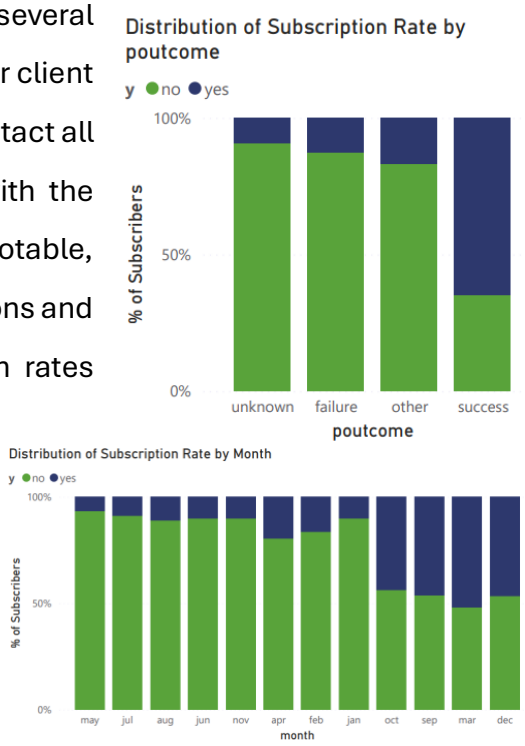


Figure 2 Client Data vs Subscription

5.3 Data Transformation

We mapped the binary variables (`housing and loan`) from numeric to categorical values. We detected outliers using the IQR method and found that features like `balance, duration, campaign, pdays, and previous` had high outlier percentages. We normalized all numerical features using Min-Max scaling to bring them into a common range. For the categorical variables, we applied one-hot encoding with '`drop_first=True`' to avoid multicollinearity. After these steps, the final dataset of 45,211 records and 40 features, optimized for memory and ready for modeling

6. Modelling

In the context of improving term deposit subscription rates at a Portuguese bank, predictive modeling plays a crucial role in identifying clients most likely to respond positively to marketing campaigns. The models help shift from random cold-calling toward a focused, data-driven targeting strategy. Four distinct models were chosen for their different strengths in prediction, interpretability, and computational complexity: Logistic Regression, Decision Tree, Random Forest, and Neural Network. Each offers unique advantages for both business interpretation and operational deployment.

6.1 Logistic Regression

Logistic Regression serves as a foundational statistical model for binary classification problems, such as predicting whether a customer will subscribe (yes/no). This model was chosen for its simplicity and interpretability. It provides a clear view of how individual predictors—such as call duration or whether the customer was previously contacted—affect the likelihood of subscription. Its coefficients, when exponentiated, can be directly interpreted as odds ratios, which is valuable for business stakeholders. With an AUC of 0.8895 and a sensitivity of 64.85%, it provided a solid baseline for comparison.

6.2 Random Forest

Random Forest is an ensemble learning technique that aggregates the output of multiple decision trees to improve prediction accuracy and reduce overfitting. It was selected for its robustness and ability to capture complex interactions between features without heavy preprocessing. Its interpretability is enhanced through measures like MeanDecreaseGini, which rank features by importance. The model achieved the highest AUC of 0.9122, with a strong

sensitivity of 74.52%, indicating its capability to detect true subscribers effectively. Random Forest balances predictive power and interpretability, making it suitable for operational deployment.

6.3 Decision Tree

The Decision Tree model was employed due to its intuitive, rule-based structure that mimics human decision-making. Although it achieved the lowest AUC of 0.75 among the models, it's clear logic and if-then rules make it extremely valuable for business interpretation and auditability. For instance, it might reveal rules such as 'if call duration is less than 90 seconds and month is May, then customer is unlikely to subscribe.' Such insights are easy to explain to non-technical stakeholders and useful for deriving business heuristics.

6.4 Neural Network

Neural Networks are powerful machine learning models that can learn complex, non-linear relationships from data. While less interpretable, they often outperform traditional methods when tuned properly. In this study, the neural network achieved an AUC of 0.91, a sensitivity of 74.27%, and accuracy of 86.88%. Its strength lies in identifying patterns that simpler models might miss—making it particularly suited for nuanced segmentation in marketing. However, the trade-off is interpretability, so it's best used in conjunction with more transparent models when actionable insights are required.

7. Model Comparison and Business Relevance:

Each model brought distinct advantages. Logistic Regression provided transparency and was easy to communicate. The Decision Tree offered explicit rules that could be directly implemented in call center scripts. Random Forest balanced accuracy and feature importance insights, making it ideal for strategic planning. Neural Networks, though a 'black box', yielded the best performance and were most suitable for automation and real-time predictions. Together, these models informed both tactical campaign design and long-term customer engagement strategies.

Model Performance: Confusion Matrices and AUC Scores:

To quantify the performance of each model, we include confusion matrices and AUC scores. These metrics provide a clearer picture of how each model distinguishes between subscribers and non-subscribers, highlighting trade-offs between sensitivity and specificity.

Model	Accuracy	Sensitivity	Specificity	AUC
Logistic Regression	87.48%	64.85%	90.47%	0.8895
Random Forest	87.4%	74.52%	89.1%	0.9122
Decision Tree	86.0%	58.4%	89.6%	0.75
Neural Network	86.88%	74.27%	88.5%	0.91

Table 3 Machine Learning Models Comparison

Evaluation:

Among the four models tested, Random Forest was selected as the final model for deployment. It offered the highest AUC (0.9122), excellent sensitivity (74.52%), and balanced specificity (89.1%), making it ideal for accurately identifying potential subscribers without over-targeting uninterested clients. Random Forest also provided insights into variable importance, which informed marketing strategies.

To assess the real-world impact of the model, a comparative evaluation was conducted. In a traditional, non-model campaign, random calling yielded a conversion rate of only 11.7%. However, when the model was used to target the top predicted customers, the conversion rate rose dramatically to 47.5%, reducing the number of calls needed per subscriber from 9 to just 2.

Marketing Efficiency Comparison:

Metric – test size – 30%	Without Model (Random Calling)	With Model (Random Forest)
Total Customers Contacted	13,563	2,481 (Predicted yes)
Subscriptions Gained	1,587	1,179
Conversion Rate	11.7%	47.52%
Calls per Subscriber	9	2
Non-Subscribers Contacted	11,976	1,302
Subscribers Missed	0	408
Efficiency Ratio (Subs/Non-Subs)	13%	91%

Table 4 Without Model vs With Model

Conclusion:

The Random Forest model demonstrated superior performance in both statistical evaluation and business impact. By accurately identifying high-probability subscribers, it enabled a more focused and efficient marketing strategy. This not only reduced operational costs by minimizing wasted calls but also improved customer experience by limiting irrelevant outreach. Based on these outcomes, Random Forest is recommended for ongoing and future predictive marketing efforts.

8. Works Cited

[1]Berry, M. J., & Linoff, G. S. (2004). *Data Mining Techniques - Second Edition*. Indianapolis, Indiana: Wiley Publishing, Inc. Retrieved from Wiley Publication.

[2]Chen, J. (2024, 06 10). *investopedia terms*. Retrieved from investopedia: https://www.investopedia.com/terms/t/termdeposit.asp

[3]Domingos, P. (2012, October). *A few useful things to know about machine learning*. *Commun*. Retrieved from ACM Digital Library: https://doi.org/10.1145/2347736.2347755

[4]Han, J., Kamber, M., & Pei, J. (2012). *Data Mining: Getting to know your data*. Retrieved from Science Direct: https://doi.org/10.1016/B978-0-12-381479-1.00002-2

[5]Moro, S., Cortez, P., & Rita, P. (2012, 02 13). *UCI archives*. Retrieved from UCI machine learning repository: https://archive.ics.uci.edu/dataset/222/bank+marketing

[6]Moro, S., Cortez, P., & Rita, P. (2014, 06). A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 22-31. Retrieved from ScienceDirect: https://doi.org/10.1016/j.dss.2014.03.001

[7]Shearer, C. (2000). *Data Warehousing Institute*. Retrieved from Journal Of Data Warehousing: https://mineracaodedados.wordpress.com/wp-content/uploads/2012/04/the-crisp-dm-model-the-new-blueprint-for-data-mining-shearer-colin.pdf

9. Appendix

```
predicted.test.class = ifelse(predicted.test.prob > 0.2, "yes", "no")
```

Model Peformance - Confusion Matrix

```
confusionMatrix(as.factor(predicted.test.class), test_data2$y, positive = "yes")
```

```
## Confusion Matrix and Statistics
##
##          Reference
## Prediction    no   yes
##        no  10814   556
##        yes  1138  1026
##
##               Accuracy : 0.8748
##                 95% CI : (0.8691, 0.8804)
##    No Information Rate : 0.8831
##    P-Value [Acc > NIR] : 0.9986
##
##                  Kappa : 0.4772
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.64855
##             Specificity : 0.90479
##          Pos Pred Value : 0.47412
##          Neg Pred Value : 0.95110
##              Prevalence : 0.11689
##          Detection Rate : 0.07581
##    Detection Prevalence : 0.15989
##       Balanced Accuracy : 0.77667
##
##        'Positive' Class : yes
##
```

Logistic Regression Model – Performance

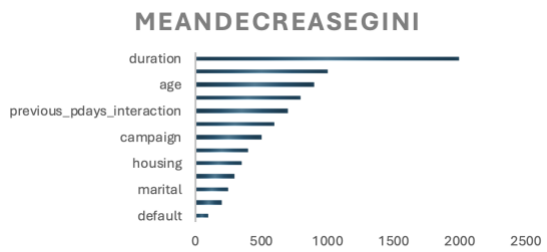
```
# Apply cutoff of 0.2
pred.labels <- ifelse(pred.prob > 0.22, "yes", "no")

#Convert predictions to factor with same levels as the actual data
pred.labels <- factor(pred.labels, levels = c("no", "yes"))


# Confusion matrix using the new predictions
conf_matrix <- confusionMatrix(pred.labels, test_data2$y, positive = "yes")
print(conf_matrix)
```

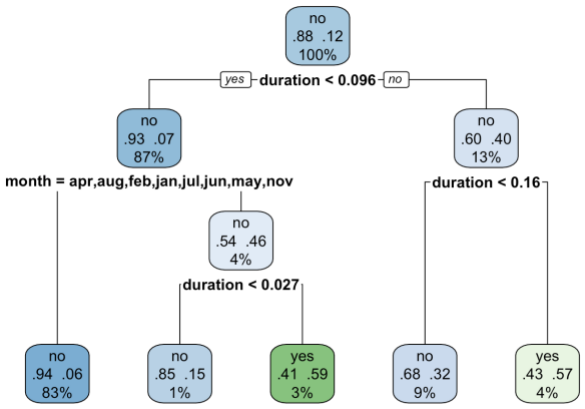
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    no   yes
##        no  10650   403
##        yes  1302  1179
##
##               Accuracy : 0.874
##                 95% CI : (0.8683, 0.8796)
##    No Information Rate : 0.8831
##    P-Value [Acc > NIR] : 0.9995
##
##                  Kappa : 0.5105
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.74526
##             Specificity : 0.89106
##          Pos Pred Value : 0.47521
##          Neg Pred Value : 0.96354
##              Prevalence : 0.11689
##          Detection Rate : 0.08711
##    Detection Prevalence : 0.18332
##       Balanced Accuracy : 0.81816
##
##        'Positive' Class : yes
##
```

Random Forest Model – Performance



Feature	no	yes	MeanDecreaseAccuracy	MeanDecreaseGini
duration	52.8601276	86.888975	85.8074664	1941.7284
housing	26.2806398	20.503937	34.8640701	188.39321
month	30.5203608	17.0952675	32.9912795	665.06003
previous_pdays_interaction	21.6330531	25.635425	27.8608753	469.76915
contact	22.6649353	9.9032561	24.2748041	166.97965

Random Forest – Variable Importance



Decision Tree

```r
# Apply cutoff of 0.2 to classify as "yes" or "no"
pred.labels <- ifelse(pred.prob > 0.2, "yes", "no")
pred.labels <- factor(pred.labels, levels = c("no", "yes"))

# Confusion matrix using custom cutoff
conf_matrix <- confusionMatrix(pred.labels, test_data2$y, positive = "yes")
print(conf_matrix)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    no   yes
##        no  10716   657
##        yes  1236   925
##
##               Accuracy : 0.8601
##                 95% CI : (0.8542, 0.8659)
##    No Information Rate : 0.8831
##    P-Value [Acc > NIR] : 1
##
##                  Kappa : 0.4153
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.58470
##             Specificity : 0.89659
##          Pos Pred Value : 0.42804
##          Neg Pred Value : 0.94223
##              Prevalence : 0.11689
##          Detection Rate : 0.06835
##    Detection Prevalence : 0.15967
##       Balanced Accuracy : 0.74064
##
##        'Positive' Class : yes
```

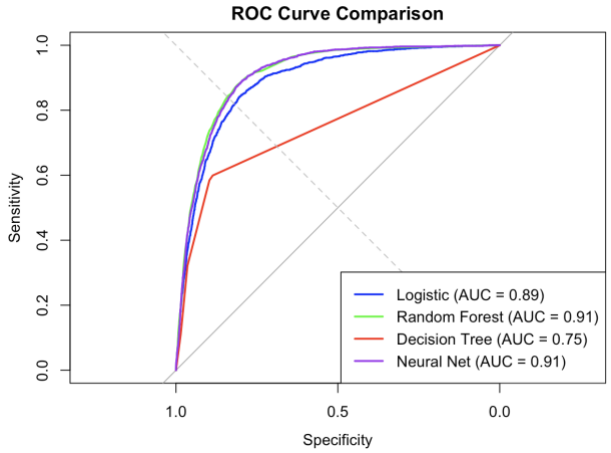
Decision Tree – Model Performance

```r
# For binary classification:
nn_preds <- predict(nn_model, newdata = test_data2, type = "class")
```

```r
nn_preds <- factor(nn_preds)
confusionMatrix(nn_preds, test_data2$y, positive = "yes")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    no   yes
##        no  11515   910
##        yes   437   672
##
##               Accuracy : 0.9005
##                 95% CI : (0.8953, 0.9055)
##    No Information Rate : 0.8831
##    P-Value [Acc > NIR] : 7.074e-11
##
##                  Kappa : 0.4461
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.42478
##             Specificity : 0.96344
##          Pos Pred Value : 0.60595
##          Neg Pred Value : 0.92676
##              Prevalence : 0.11689
##          Detection Rate : 0.04965
##    Detection Prevalence : 0.08194
##       Balanced Accuracy : 0.69411
##
##        'Positive' Class : yes
```

Neural Network Model – Performance



Models Area Under the Curve - Comparison