Portuguese Banking Subscription Predictions

Andrew Wessel
May 21st 2022

Problem outline

Term Deposits are fixed-term investments where a client deposits money into an account at a bank for a given time period (e.g. 1 month to >5 years)

Banks loan out the money and collect interest, some which is paid back to clients and some remains as profit

Banks have two major ways to increase profits from term deposits

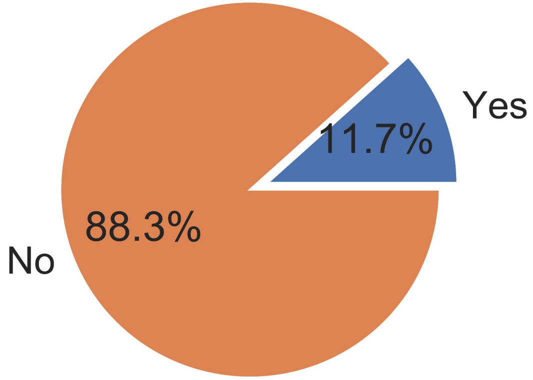
- Increasing the amount of money invested per term deposit
- Increasing the number of clients to subscribe to term deposits

Is there a way to predict clients who are more likely to subscribe to term deposits?

Project & Data

- Two fold project
  - Develop model to predict probability of customer subscribing to term deposit
  - Determine maximum profit for subscribing customers to the term deposit
- Dataset
  - 45,211 bank clients' anonymized demographic, financial, and bank history data from Portuguese banks from 2008-2010
  - Information about current and previous campaign history
  - Output variable: whether client has subscribed to a term deposit (yes/no)

Distribution of Term Deposits



https://archive.ics.uci.edu/ml/datasets/bank+marketing
S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

Exploratory Data Analysis

Input variables:

1. *age* (numeric)
2. *job* : type of job (categorical)
3. *marital* : marital status (categorical)
4. *education* (categorical)
5. *default*: has credit in default? (binary)
6. *balance*: average yearly balance, in euros (numeric)
7. *housing*: has housing loan? (binary)
8. *loan*: has personal loan? (binary)
9. *contact*: contact communication type (categorical)
10. *day*: last contact day of the month (numeric)
11. *month*: last contact month of year (categorical)
12. *duration*: last contact duration, in seconds (numeric)
13. *campaign*: number of contacts performed during this campaign and for this client (numeric, includes last contact)
14. *pdays*: number of days that passed by after the client was last contacted from a previous campaign (numeric)
15. *previous*: number of contacts performed before this campaign and for this client (numeric)
16. *poutcome*: outcome of the previous marketing campaign (categorical)

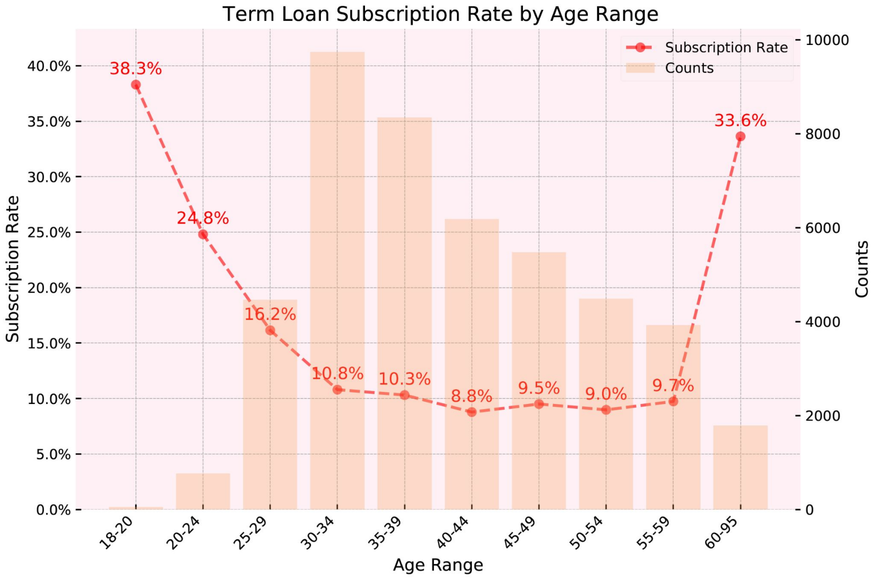
Input Variable Data Types:

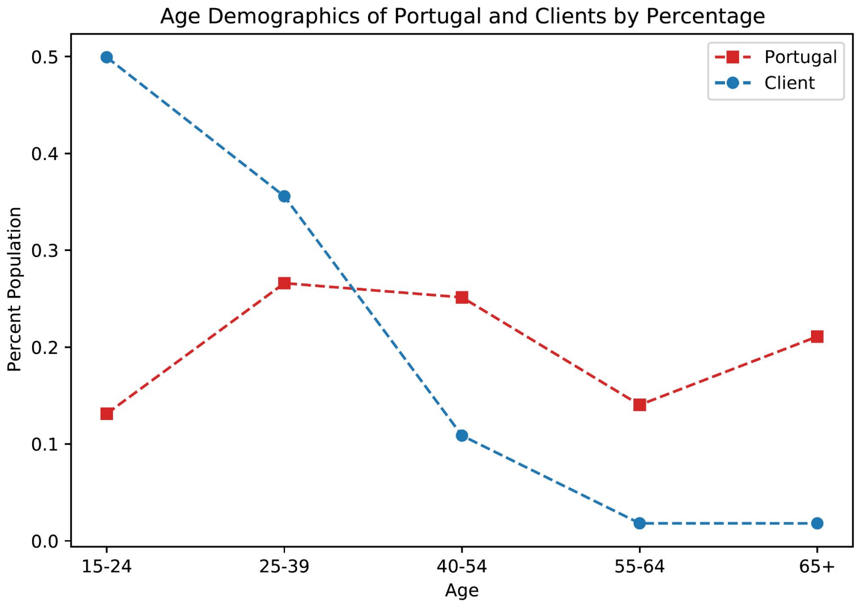
- bank client data
- related with the last contact of the current campaign
- other attributes

Output variable (desired target):

17. *y* - has the client subscribed a term deposit? (binary: "yes","no")

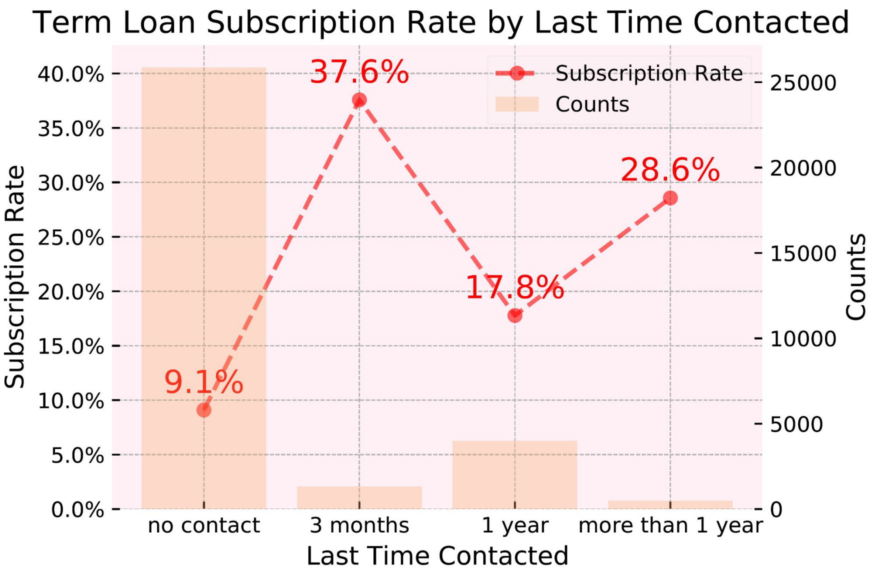
Exploratory Data Analysis - Age

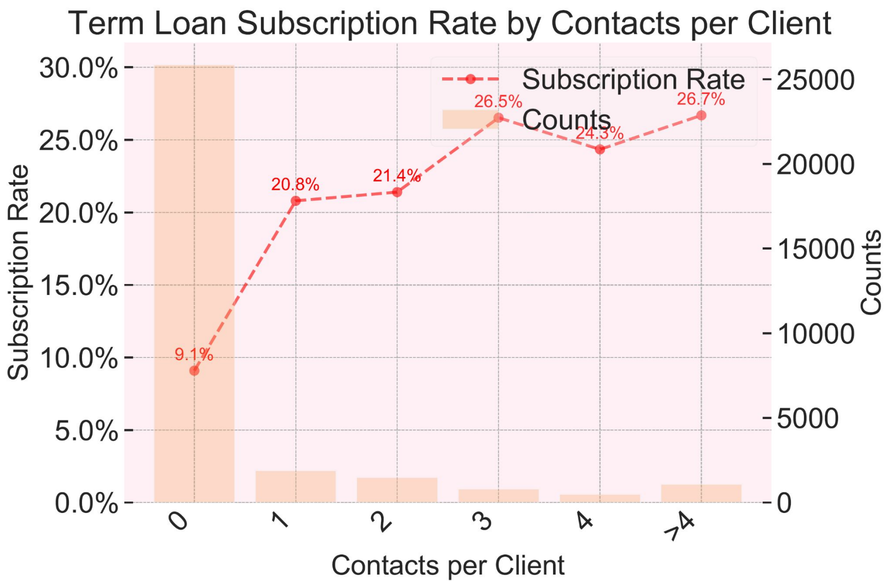




- 85% of clients are under the age of 40
- Students (< 25 years old) and retired (> 60 years old) explain deviations in Subscription Rate (~10%)

Exploratory Data Analysis - Call Frequency

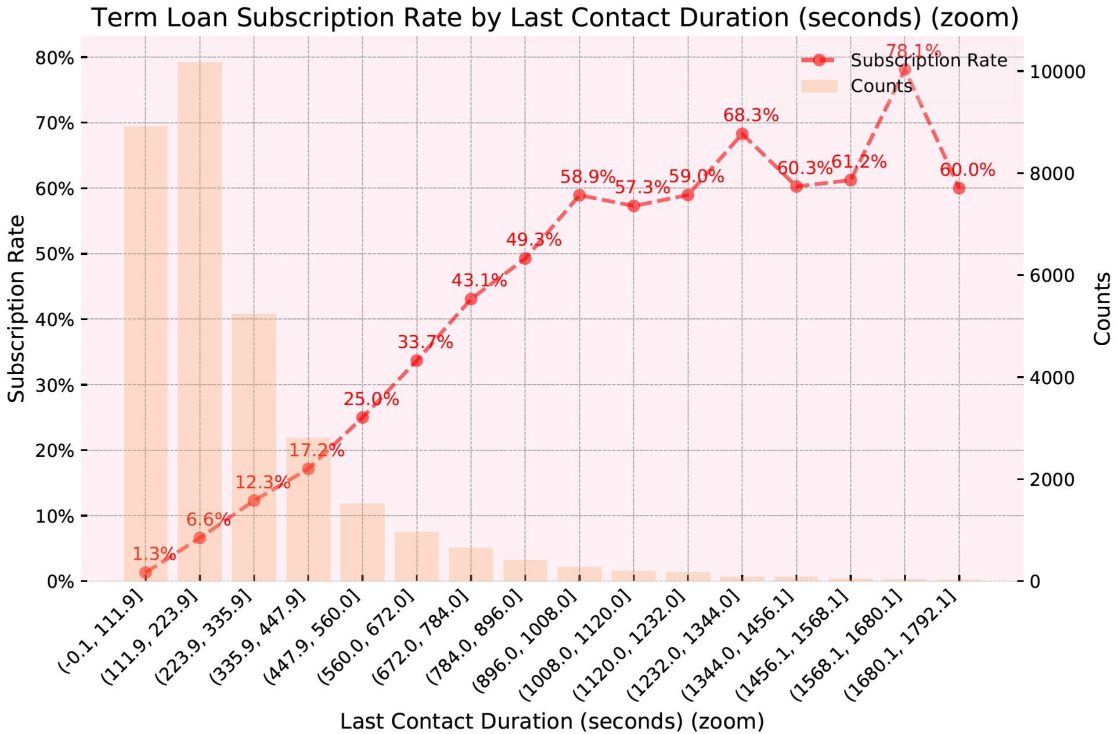




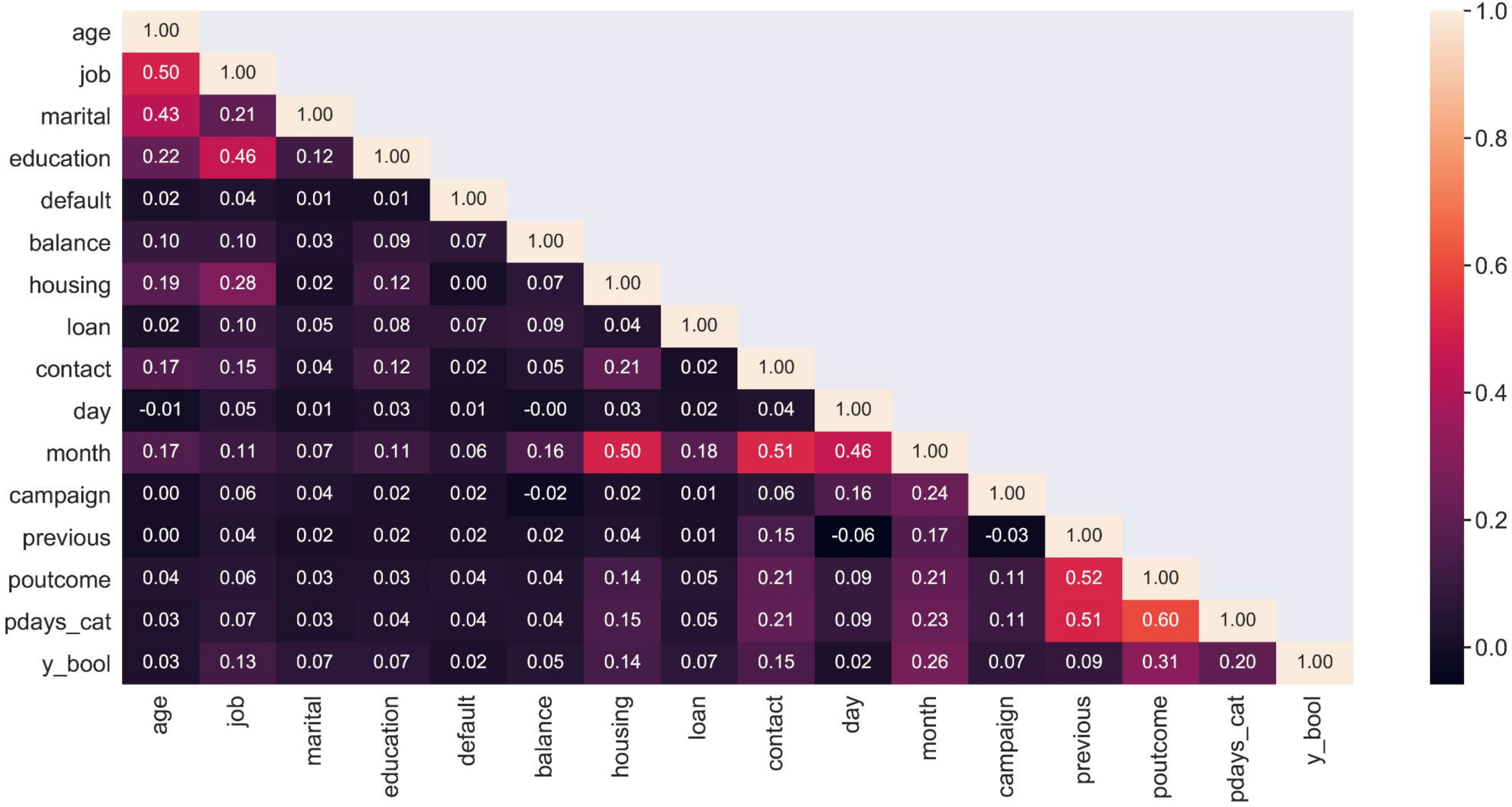
- Previous contacts for clients roughly doubles subscription rate
- Clients previously contacted at least once roughly doubles subscription rate

Exploratory Data Analysis - Call Duration

- Call duration <u>cannot</u> be used in modeling
  - If client was not contacted before, then output value will be “no”
  - Duration of call cannot be known prior to the solicitation
  - Should be considered only as benchmark
- Duration <u>can</u> be used to help develop marketing strategies
  - Most calls lasted 5-6 minutes or less
  - Subscription rate increased up until 16 minutes, then remains constant



Exploratory Data Analysis - Heat Map



Heat Map of Features Generated from Cramer's V, Pearson’s R, and Correlation Ratio

Modeling

- Three classification models were chosen
  - Random Forest
  - Logistic Regression
  - XGBoost
- Hyperparameters were chosen using GridSearch
- ROC-AUC scores for all were similar at 0.80
- Each model produced roughly a 66% accuracy rate

Classifier	ROC-AUC Score	Best Hyperparameter Values
Random Forest	0.797	{'bootstrap': True, 'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 10, 'n_estimators': 275}
Logistic Regression	0.782	{'C': 75.43, 'penalty': 'L2''}
XGBoost	0.804	{'objective': 'binary:logistic', 'learning_rate': 0.10, 'max_depth': 4}

Performance for the Random Forest, Logistic Regression, and XGBoost Classifiers Note: ROC-AUC Scores were generated on test data.

Analysis Result - Profit Curve Assumptions

- Term Deposit Assumptions:
  - 1 year, €10,000 term deposit at 1% return
  - Interest compounded daily
  - Total revenue for term deposit: €158.36
- Operation Cost Assumptions:
  - Monthly salary: € 3,000
  - 20 days/month, 8 hour/day
  - Cost per hour: € 18.75

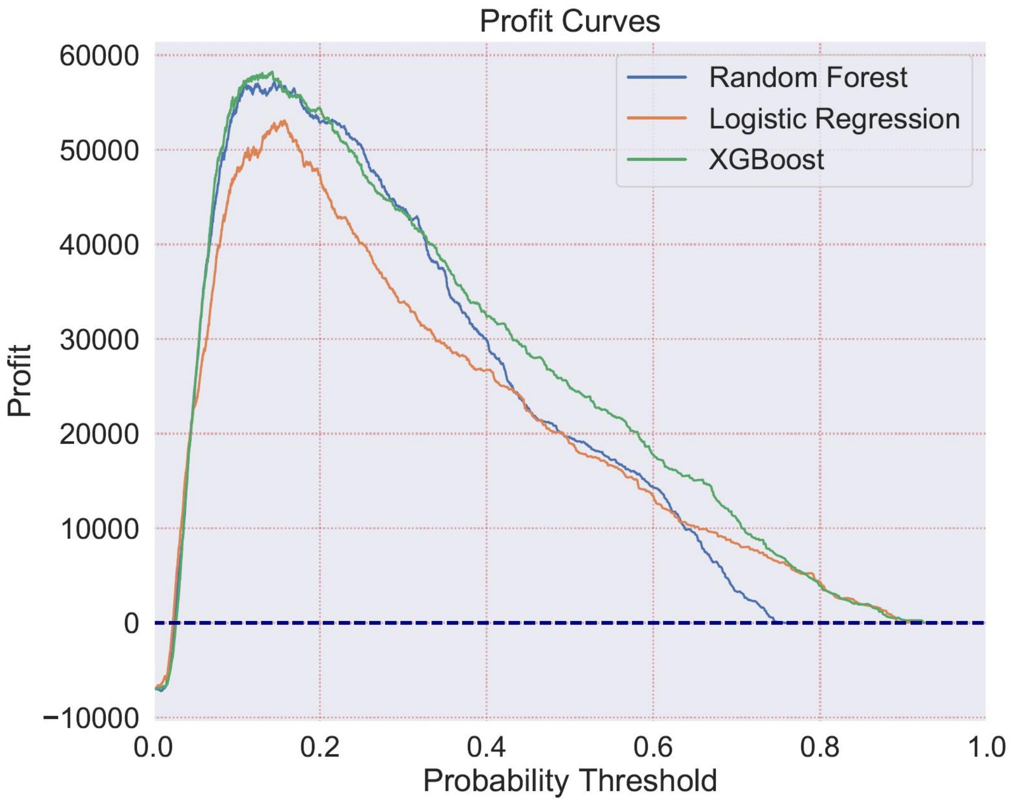
	Predicted Success	Predicted Failure
Actual Success	**TP** €139.61	**FN** €0
Actual Failure	**FP** − €18.75	**TN** €0

Cost Matrix for Profit Curve

https://tradingeconomics.com/portugal/lending-interest-rate-percent-wb-data.html
https://www.frbsf.org/economic-research/wp-content/uploads/sites/4/Session-1-Paper-2-Savov.pdf
https://fred.stlouisfed.org/series/DDEI01PTA156NWDB
https://www.cgd.pt/Particulares/Poupanca-Investimento/Depositos-a-Prazo-e-Poupanca/Pages/Deposito-Prazo-3-Anos.aspx

Analysis Result - Profit Curves

- XGBoost had the highest profit but all models had similar results
- Maximum Profit of €58,000 at probability threshold of 14.2%



Recommendations for Subscription Success

- Contacting clients
  - Average client contacted 3 times increases odds of success rather than no contact
  - Contact within last 3 months has highest increase in subscription rate but effects seem to last
  - Keeping clients on the call and engaged had higher subscription success rate
- Most clients are younger than 40 years old
  - Marketing to this demographic should increase success
- Profit Analysis
  - Subscription target of 14% of clients to maximize profits

Considerations, Improvement, and Future Work

- The models’ prediction rate was 80%
  - Further optimization steps can be taken
  - Additional models could be attempted to improve viability
- External market factors were in play during data window of 2008-2010
  - Additional client data for windows before and after global financial crisis
  - Austerity measures were in place during this window, clients may be less likely to subscribe
  - Inflation may affect dollar amounts for profits over larger windows
- Incorporating additional demographic data for Portugal and comparing to client subset
  - Age analysis saw that most clients are younger, unlike the fairly distributed values for Portugal
  - Job, education, marital status, etc.information could hone in on new marketing strategies