


# Banking Dataset

# Marketing Targets Prediction

Customer conversion prediction for term deposits

Group 8

# Background and Motivation

- Every business has a limited marketing budget. Therefore it's vital that each dollar is spent in the most efficient way possible.
  - Prediction based on existing customer base really helps in developing focussed marketing initiative and better customer upselling
  - Product: Term deposits are similar to GICs, where a client will give the bank money in exchange for the money plus interest after a fixed period of time. During this time, the client is unable to withdraw their money.
  - In a bank's case, it's necessary to determine which of their clients will be receptive to phone marketing campaigns regarding the bank's financial services, specifically term deposits
- 

# Data Source

- Dataset source: Kaggle  
(<https://www.kaggle.com/prakharrathi25/banking-dataset-marketing-targets>)
- The data was pulled from the UCI Machine learning repository. The data was gathered from marketing campaigns a Portuguese banking institution implemented through phone calls.



# Questions to answer

- Determine whether or not a bank client would be interested in a term deposit subscription based on their profile and past history with the bank. This will enable the bank to better target their phone based marketing efforts towards clients who would be open to a term deposit subscription.
- Determine if a relationship between a client's profile and their usage of financial services exists



# Tools & Technology

- Data Cleaning and Analysis: The Python "Pandas" and "SciKitLearn" machine learning library used for data preprocessing (eg. clean the data) and perform an exploratory analysis. Deep analysis conducted using Python.
- Database Storage: Postgres is the database used since the data for the topic is well structured with a specific schema. Data imported and exported data using SQL queries and the Python "Pandas" library.
- Dashboard : Js and Plotly used for a fully functioning and interactive dashboard and hosted on Github Pages.



# Preprocessing

- One hot encode categorical features (job, marital, education, contact, poutcome, month)
  - Ensured that analysis and model can work with categorical data
- Dropped features related to last contact of current campaign (day, month, duration)
  - When conducting a future marketing campaign, these would be chosen by the campaign so shouldn't be a factor in determining who to contact about a term deposit
- Data split into stratified train and test sets (80/20)
  - More data for training compared to the default 75/25 split to make up for dataset size
  - Stratified to handle class imbalance
- Upsample with SMOTE
  - Deal of class imbalance of the outcome variable "y"



# Data exploration

**balance:** Bank account balance

- Outliers on the negative and positive end

**pdays:** Past campaign contact

- Majority of participants weren't a part of a previous campaign

**previous:** Number of times contacted before campaign

- Majority of participants haven't had phone contact with the bank prior to marketing campaign

	age	balance	day	duration	campaign	pdays	previous
count	49732.000000	49732.000000	49732.000000	49732.000000	49732.000000	49732.000000	49732.000000
mean	40.957472	1367.761562	15.816315	258.690179	2.766549	40.158630	0.576892
std	10.615008	3041.608766	8.315680	257.743149	3.099075	100.127123	2.254838
min	18.000000	-8019.000000	1.000000	0.000000	1.000000	-1.000000	0.000000
25%	33.000000	72.000000	8.000000	103.000000	1.000000	-1.000000	0.000000
50%	39.000000	448.000000	16.000000	180.000000	2.000000	-1.000000	0.000000
75%	48.000000	1431.000000	21.000000	320.000000	3.000000	-1.000000	0.000000
max	95.000000	102127.000000	31.000000	4918.000000	63.000000	871.000000	275.000000

# Data exploration: Feature Distributions

**age:** Age of participant

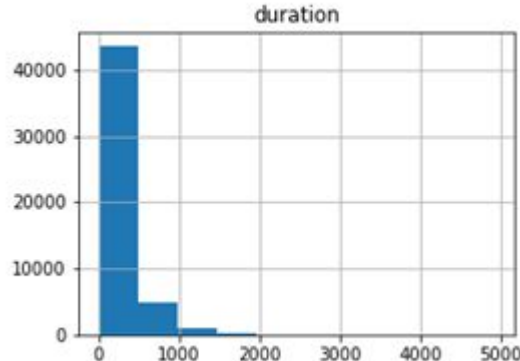
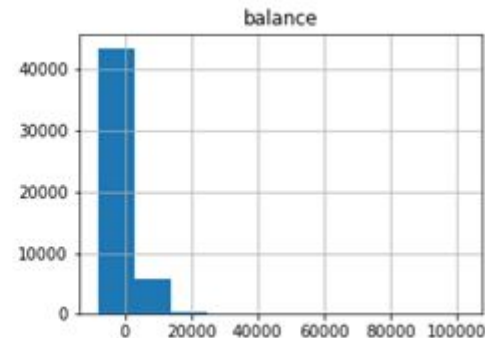
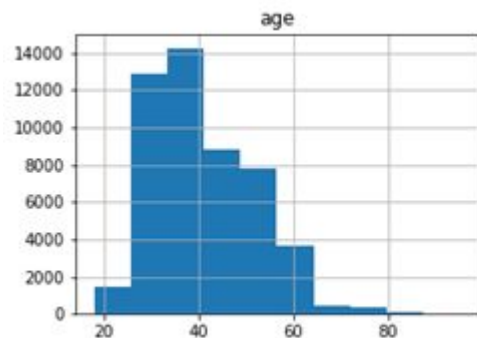
- Most participants are between 20-50 years old

**balance:** Bank account balance

- Most people have less than \$1500 in their bank account

**duration:** Seconds before end of day when contacted during campaign

- Majority of contacts were around the same time of day





# Data exploration: Feature Distributions

**month:** Month when contacted during campaign

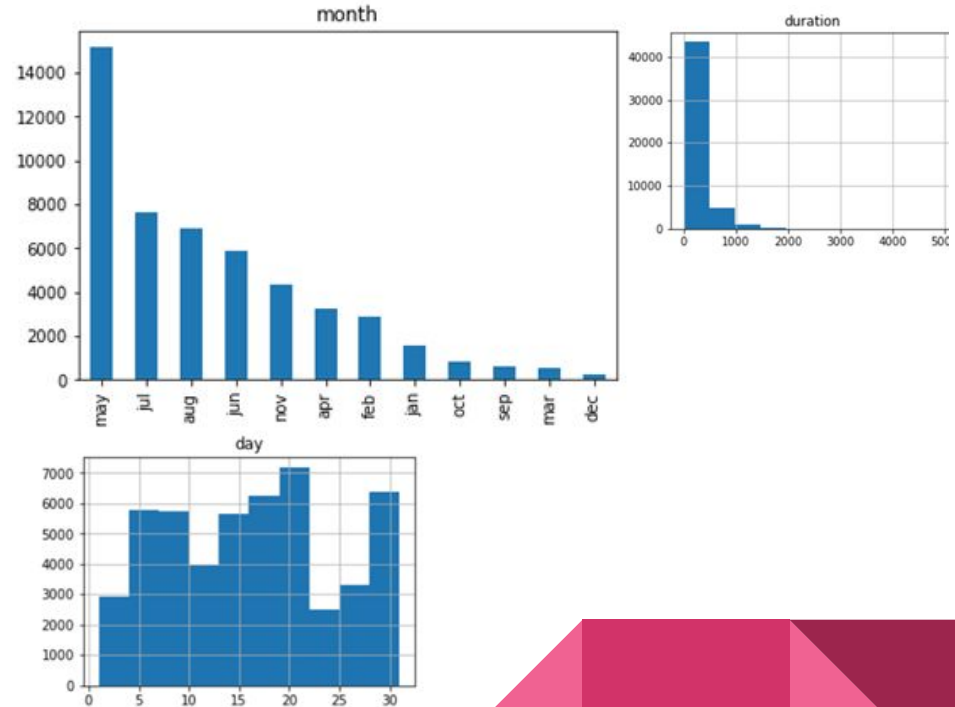
- Campaign was held primarily during the summer

**day:** Day of month when contacted during campaign

- No specific day was favored

**duration:** Seconds before end of day when contacted during campaign

- Majority of contacts were around the same time of day



# Data exploration: Feature Distributions

**default:** User has credit in default

- Almost no one has defaulted

**housing:** Has a housing loan

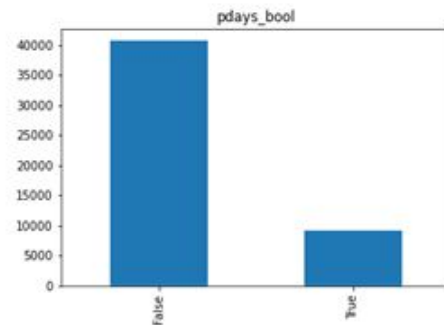
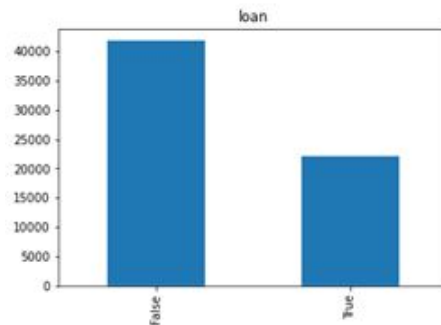
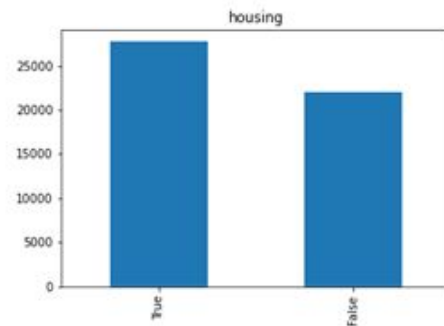
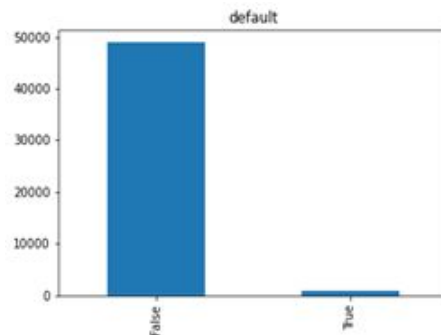
- Little more than half of the participants have a housing loan

**loan:** Has a personal loan

- 20% of participants have a personal loan

**pdays\_bool:** Whether the participant was contacted in a past campaign

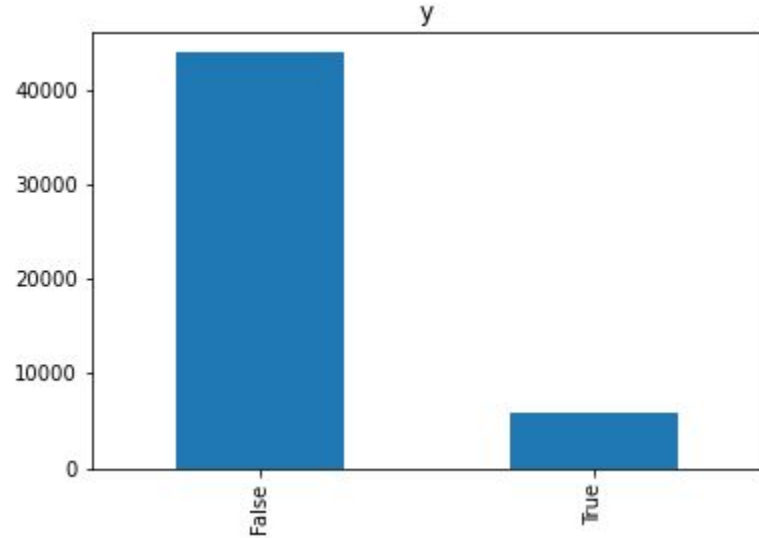
- Most people were not contacted in a past campaign



# Data exploration: Feature Distributions

**y:** Client has subscribed to a term deposit

- A class imbalance between True and False is present



## Data exploration: Correlations

**y:** Whether the participant has subscribed to a term deposit

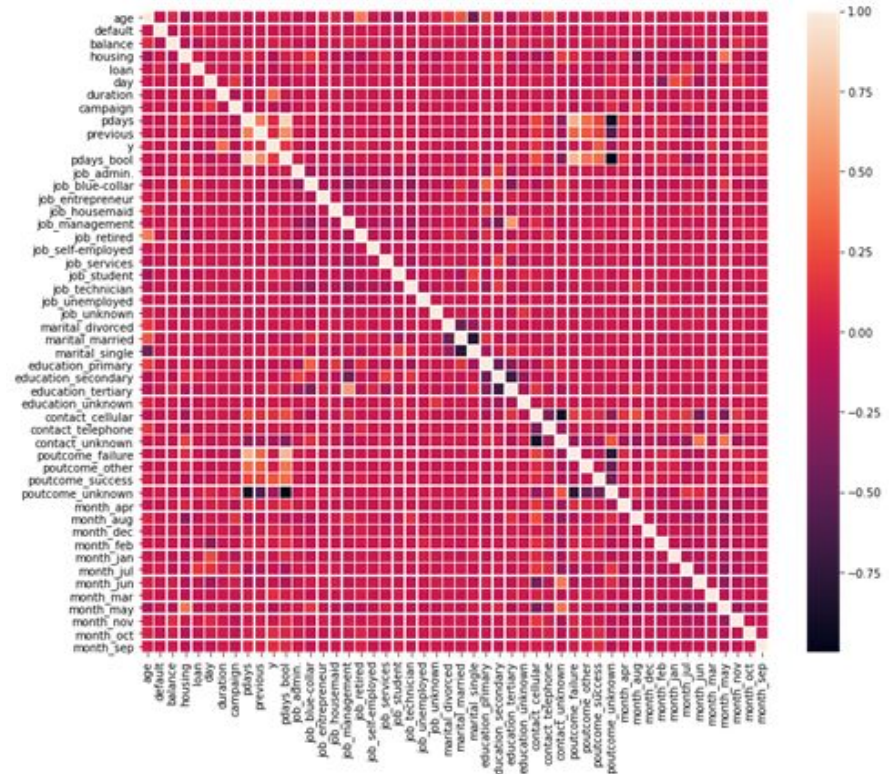
- Y is not correlated strongly with any particular variable. The strongest correlation is with housing

**age:** Age of participant

- Age is correlated with each of the following features: housing, being married and being a student

**education:** Education of the participant

- Correlated with job



# Data exploration: Feature Importance

**job, education, marital:** characteristics about the participant

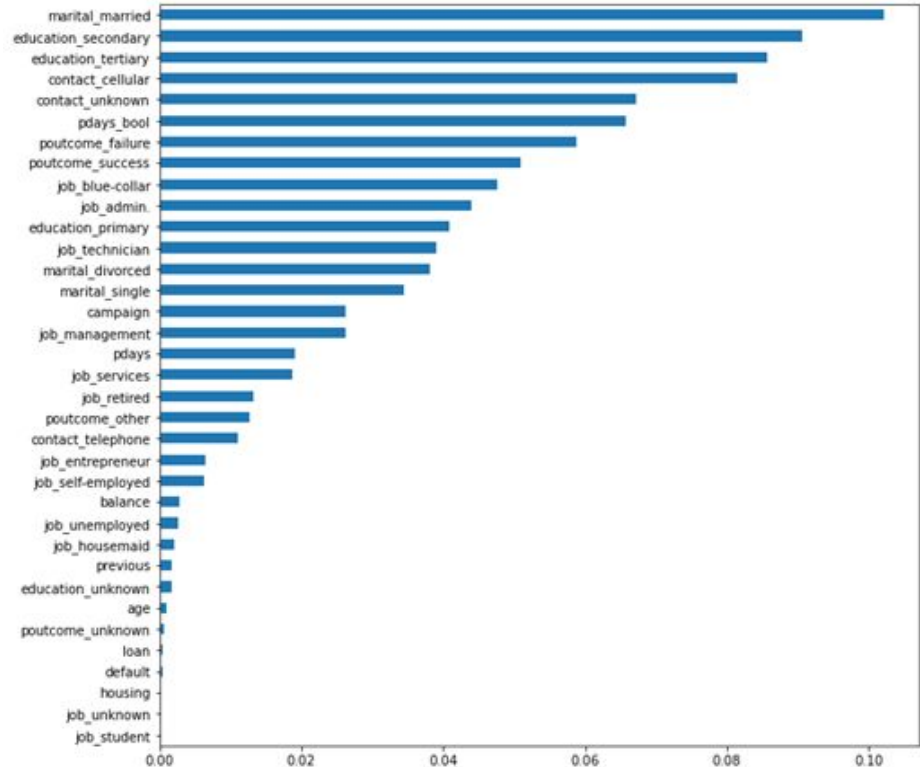
- All 3 play a strong role in determining outcome

**poutcome:** Outcome of past marketing campaign

- Strong importance in outcome

**loan, housing, age :** Whether or not participant has a personal or housing loan, age

- Surprising little importance in outcome



# Results: Accuracy and Confusion Matrix

## Accuracy

- Training and test set have a similar accuracy indicating little to no overfitting

Training Set	88.6%
Test Set	88.5%


## Confusion Matrix

- False positives exceed true positives
  - Wasted calling
- False negatives exceed true positives
  - Missing opportunity

	Predicted False	Predicted True
Actual False	8528	257
Actual True	886	276

# Results: Classification Report

	Precision	Recall	Specificity	F1 score	Geometric mean	Index balanced accuracy	Support
<b>Actual: False</b>	91%	97%	24%	94%	48%	25%	8785
<b>Actual: True</b>	52%	24%	97%	33%	48%	21%	1162
<b>Avg/Total</b>	86%	87%	37%	86%	53%	29%	9947

- Actual True: poor recall & precision performance
  - Actual False: strong recall & precision performance
- 



# Further Recommendations

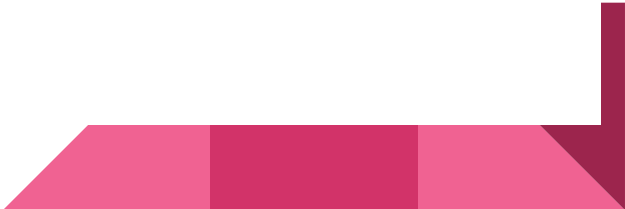
## **Cost of False Positives**

- Call center cost and member abrasion cost
  - Profitability of marketing to participants identified by model

## **A/B testing marketing outreach**

- Determine the optimal script for marketing Term Deposits

## **Additional features**

- Location of participants
  - Transaction summary/history
- 



# Project improvements

## Advance modelling techniques

- Research what modern data science techniques and models are used in marketing
- Use more model stacking and ensemble methods to improve performance metrics

