Leveraging customer information for strategic telemarketing in the banking industry

Author: Isaac Ghebregziabher

Capstone Project two

September 16, 2021

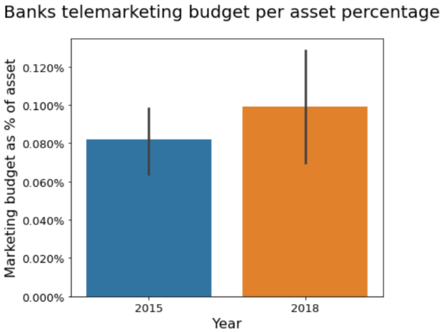
Springboard

What is telemarketing?

Telemarketing is a method of direct marketing in which a salesperson solicits prospective customers to buy products or services, either over the phone or through a subsequent face to face or web conferencing appointment scheduled during the call.

Telemarketing cost has been increasing in the banking industry

- Bank marketing cost increased over two years
- Cost as high as 0.15% of total bank's asset



Predictive modeling could increase marketing success

Object and project deliverable

Deliverable: ML model for predicting whether a customer will subscribe to a term-deposit

> *A term deposit is a fixed-term investment that includes the deposit of money into an account at a financial institution*

- The developed model will help the bank:
  - Cluster its customers into meaningful groups
  - Predict customer response to its telemarketing campaigns
  - Identify target customer groups for its future tele-marketing campaigns

Stakeholders and telemarketing data

– Our client is a Portuguese banking institution.
– Brought to us data related to telemarketing campaign
– The data consists of the following details:

- **Demographics** (age, job, education, marital status),
- **Financial data** (credit, housing loan, personal loan),
- **Contact details** (such as method of contact and month)
- **Previous campaign data** (such as outcome of previous campaign)

Overview of the Dataset

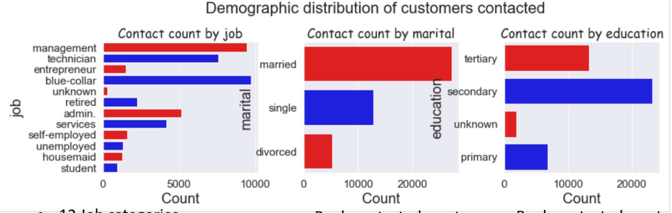
- 45211 rows X 17 columns
- 7 integer and 10 categorical type features
- No duplicates and missing values

Preview of the telemarketing dataset

	Demographics				Financial				Past an current Campaign								Campaign outcome
	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	Target
0	58	management	married	tertiary	no	2143	yes	no	unknown	5	may	261	1	-1	0	unknown	no
1	44	technician	single	secondary	no	29	yes	no	unknown	5	may	151	1	-1	0	unknown	no
2	33	entrepreneur	married	secondary	no	2	yes	yes	unknown	5	may	76	1	-1	0	unknown	no
3	47	blue-collar	married	unknown	no	1506	yes	no	unknown	5	may	92	1	-1	0	unknown	no
4	33	unknown	single	unknown	no	1	no	no	unknown	5	may	198	1	-1	0	unknown	no

Categorical feature exploration – Demographic segmentation

- 12 job categories
- 3 marital status groups
- 4 educational levels including 1 unknown

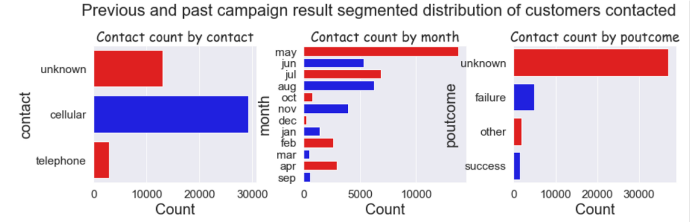


- 12 Job categories
- Highest frequency
  - Blue collar, Management
- Lowest frequency
  - Student, Housemaid

- Bank contacted most
  - Married people
- Least contacted are
  - Divorced people

- Bank contacted most
  - Secondary level
  - Tertiary
- Least contacted are
  - Primary and level

Categorical feature exploration – past and current campaign details



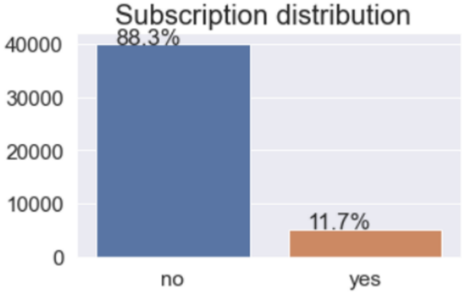
- Most contacted
    - Cellular phones
- Negligible contact
    - Land line
- Good number
    - Uknown method

- Majority contact
    - In May
- Least contact
    - December
- Average contacted
    - June, July, August

- Bank contacted most
    - Unknown past outcome
- Least contacted are
    - Success in past campaign
- Issue:
    - Other and unknown values

Target exploration – class imbalance

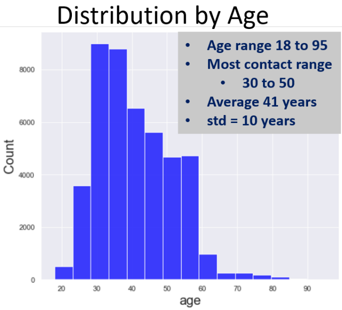
	Count: Subsciption	% (Subscription)
no	39922	88.30152
yes	5289	11.69848



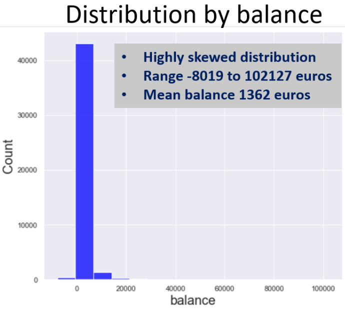
Imbalanced dataset with about 88:12 class ratio

Springboard

Range of Numerical features – Age and account balance

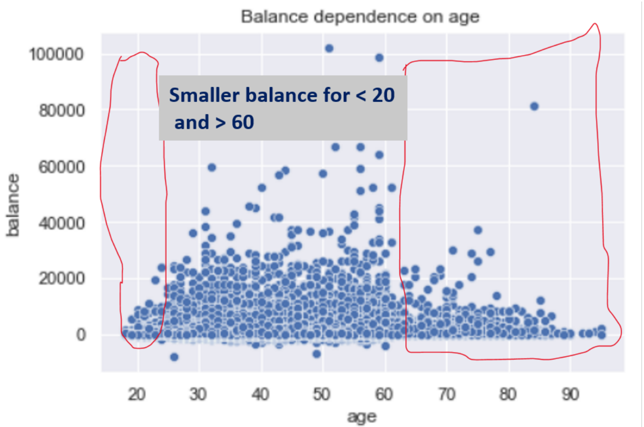


- Age range 18 to 95
- Most contact range
  - 30 to 50
- Average 41 years
- std = 10 years



- Highly skewed distribution
- Range -8019 to 102127 euros
- Mean balance 1362 euros

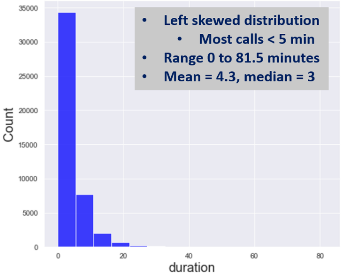
Is balance dependent on age?



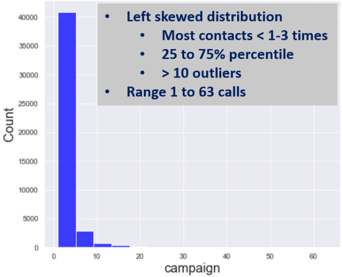
Generally balance is independent of customer's age

Range of Numerical features – Duration and number of contacts

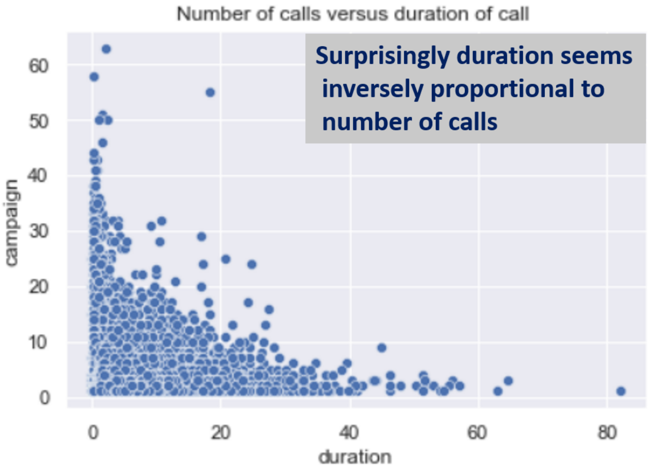
Distribution by duration



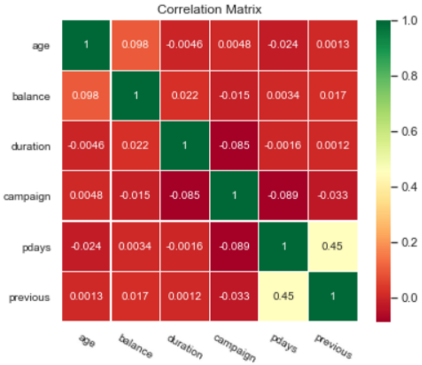
Distribution by contacts



Is duration dependent on number of contacts?



Correlation matrix and interdependence of the numerical features



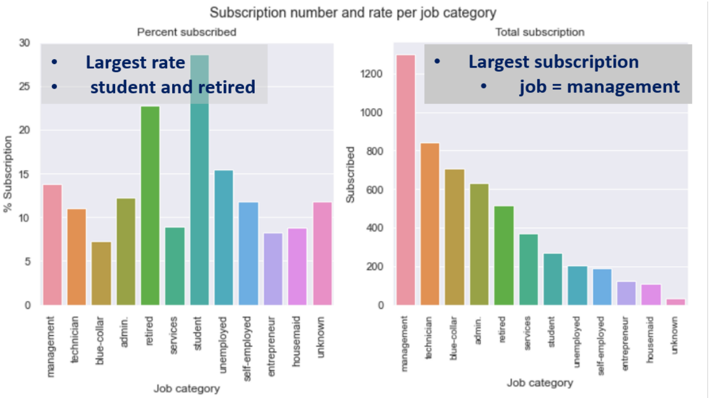
- **pdays is positively correlated with previous**
- **No significant correlation among other features.**

- **pdays** (number of days passed after previous campaign)
- **previous** (Average number of previous contacts

Objectives of Exploratory Data Analysis

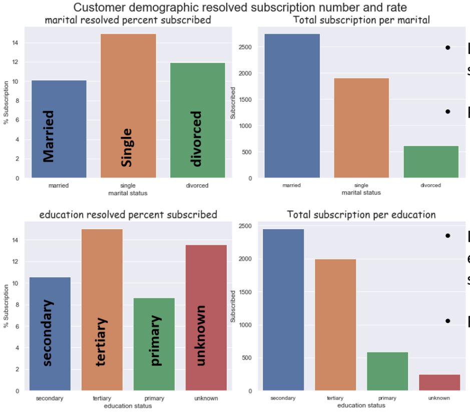
- Examine effect of each feature on target (subscription rate)
- Identify feature groups that maximize subscription rate
- Make recommendation for our client

Effect of customer job on subscription rate



- Subscription rate is the highest for students and retired individuals
- Recommendation
  - Bank should contact more of these type to maximize subscription

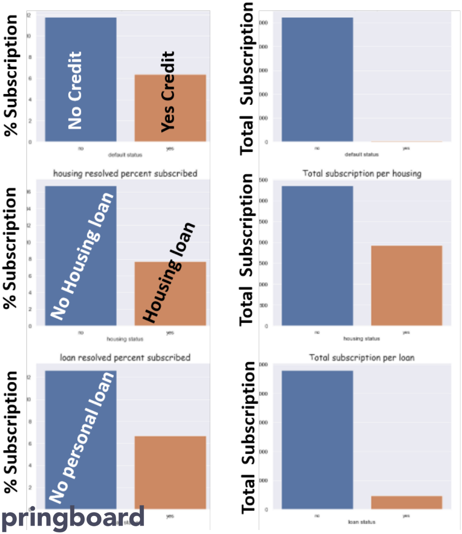
Effect of marital status and educational level on subscription rate



- Largest subscription rate for single and divorced individuals
- Recommendation
  - More singles and divorced in future campaign

- Largest subscription rate for educational level beyond high school
- Recommendation
  - More graduate students in future campaigns

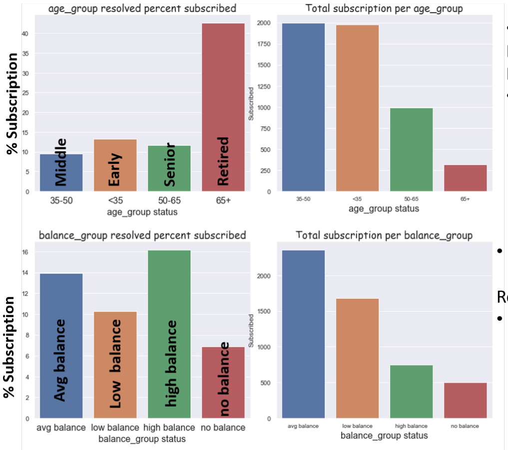
Effect of financial profile on subscription rate



- Subscription rate and number largest for customers with:

- no credit,
- no housing loan
- no personal loan

Effect of age and account balance on subscription rate



- Retired people have Highest subscription rate

Recommendation:

- Include more retired on next campaign

- People with high balance highest rate

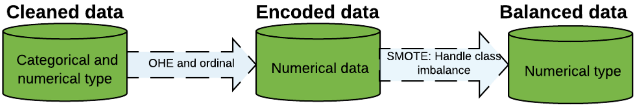
Recommendation:

- Include more high balance on next campaign

Schematics of data pre-processing

Data pre-processing included:

- Load clean dataset
- Transform categorical features
- Handle class imbalance



Transform categorical features to numerical types

Features transformed with:

- Ordinal encoding (number of labels = 2)
- One hot encoding (number of labels > 2)

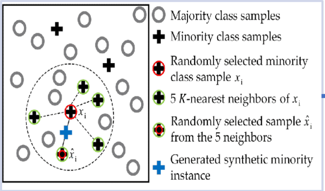
	Target	age	default	balance	housing	loan	job_blue-collar	job_entrepreneur	job_housemaid	job_management	...	job_services	job_student	job_technicia
0	0	58	0	2143	1	0	0	0	0	1	...	0	0	
1	0	44	0	29	1	0	0	0	0	0	...	0	0	
2	0	33	0	2	1	1	0	1	0	0	...	0	0	
3	0	47	0	1506	1	0	1	0	0	0	...	0	0	
4	0	33	0	1	0	0	0	0	0	0	...	0	0	

Figure: Preview of data transformed into numerical values

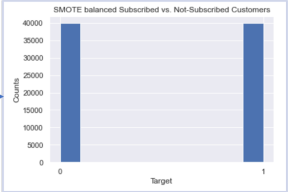
Data imbalance handled with SMOTE: Synthetic Minority Oversampling Technique

Data balanced with SMOTE shows 50:50 class ratio::

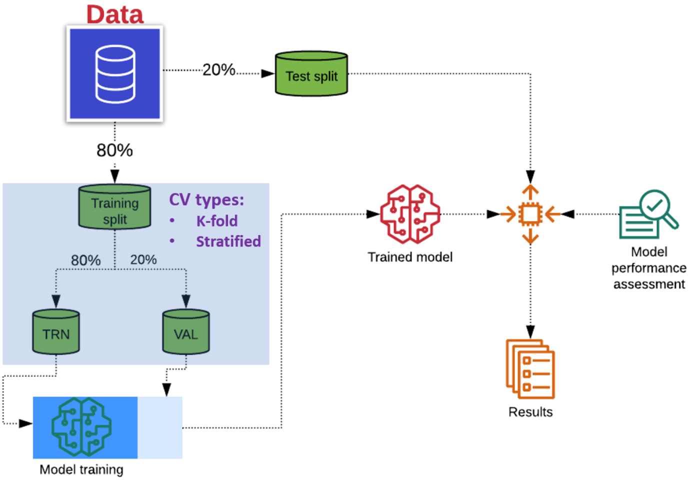
SMOTE working procedure



SMOTE balanced dataset



Schematics of machine learning modeling



Springboard

Machine learning model type corresponding to our domain problem

Supervised ML

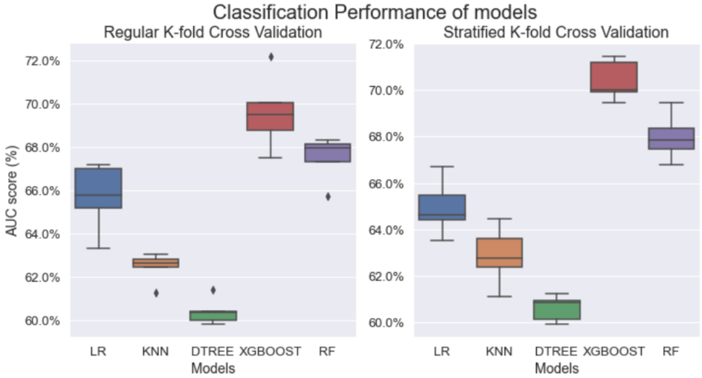
- Domain problem
- Models considered

Regression
predict: continuous

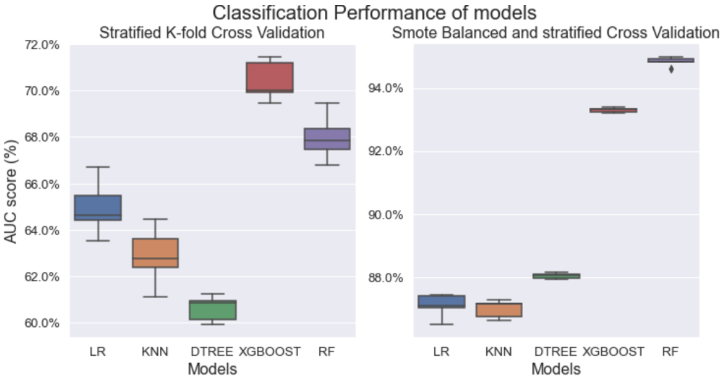
Classification
predict: yes/no class

- Logistics regression
- Decision tree
- Random forest
- XGboost
- K-nearest neighbors

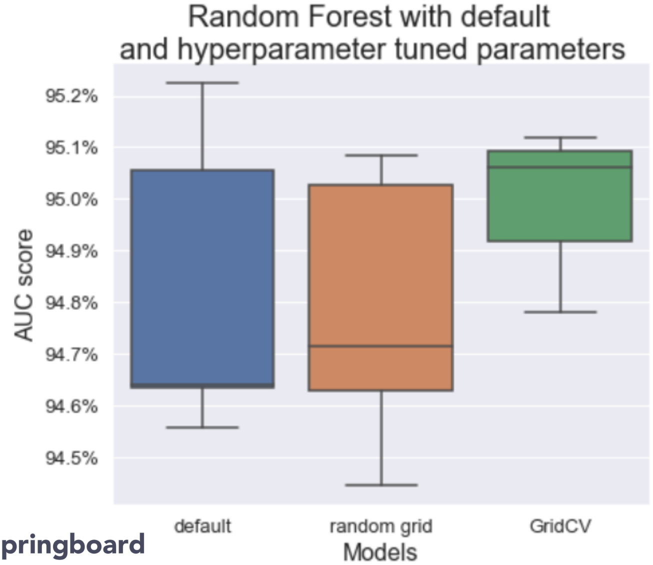
Stratified K-fold Cross validation results to marginal improvement of model performance



Data imbalance handled with SMOTE resulted to significant performance increase



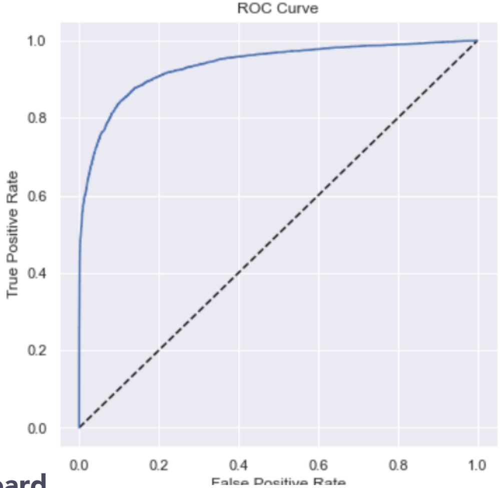
Hyperparameter tuning resulted in a stable model with slight performance improvement



Best Random Forest performance on the test set

Grid optimized Random Forest performance on the test set:

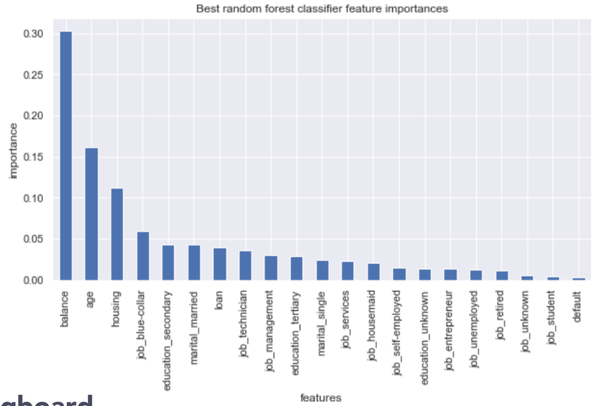
- AUC score: 95.1%



Feature importance of Best random forest model

Most important 3 features for term deposit subscription are:

- **balance**: Amount of customer account balance
- **age**: Age of customer
- **housing loan:** Has housing loan?



Summary and Future Work

- We developed supervised machine learning models for predicting customer subscription to term deposit.
- Logistic regression, decision tree, random forest, XGboos, and K-nearest neighbours machine learning algorithms considered.
- We find customer subscription to term deposit can be predicted using Random Forest with an AUC score of 95%.
- Future regression machine learning work needs to be completed to predict time spent talking to targeted customers.

Acknowledgement

Springboard mentor: Yuxuan Xin

for time generous and insightful discussions