Endeavour Group HackerRank Presentation

Case Study: Term Deposits at Lending Bank

Brendan Houng, 8 Nov 2021

Overview – Summary of Problem

Lending Bank wants to attract *term deposits* to fund its lending business. Customers receive interest on their deposits over a fixed period of time.

The bank's sales manager wants to market the product to existing clients.

Perform an analysis of the data to:

- predict - using machine learning - which existing clients are likely to subscribe to a new term deposit
- explain how different features affect the decision

- Historical information from a previous marketing campaign, including client demographics, prior call experience, market conditions

Outline

1. Problem Interpretation
2. Exploratory Data Analysis
3. Feature Engineering
4. Modelling
5. Feature Importance
6. Summary

Take you through the technical challenge:

- explaining decision making & business context
- highlight insights & aspects of the data
- add some polish to previous work (rather than large revisions)

Problem Interpretation

my goal and interpretation of the technical challenge was that I was to:

- gain an understanding of the data
- run a machine model that functions well but does not need to be ‘state of the art’ per instructions
- be able to explain the importance of the features in the model

Statistical tests, highlighting most relevant information

→

feature engineering correcting for anomalous data

→

choosing a machine learning model that is interpretable and suitable for explainability

Data Description

Column	Description
`client_id`	Unique ID of the client called [unique key]
`age_bracket`	Age bracket of the contacted client (in years)
`job`	job type of the contacted client
`marital`	marital status of the contacted client
`education`	highest level of education done by the client
`has_housing_loan`	Whether the client has a house loan (binary: yes,no)
`has_personal_loan`	Whether the client has a personal loan (binary: yes,no)
`prev_call_duration`	last contact duration (value = 0 if the client has not been contacted ever)
`days_since_last_call`	number of days that passed by after the client was last contacted from a previous campaign
`num_contacts_prev`	number of contacts performed before this campaign and for this client (numeric)
`poutcome`	outcome of the previous marketing campaign (categorical: "failure","nonexistent","success")
`contact_date`	date at which contact was made with the client (YYYY-MM-DD)
`cpi`	standing consumer price index before the call (monthly indicator)
`subs_deposit`	has the client subscribed to a term deposit? (binary: 1,0) [dependent variable]

Exploratory Data Analysis

Lending Bank’s customers from sample of 4,000 tend to be:

- 24-40 (54%), 41-60 (38.6%)
- married (59.4%)
- white collar (34%), blue collar (19%),
- Educated (31.8%)
- not have a personal loan (83.4%)
- 39.8% had a term deposit

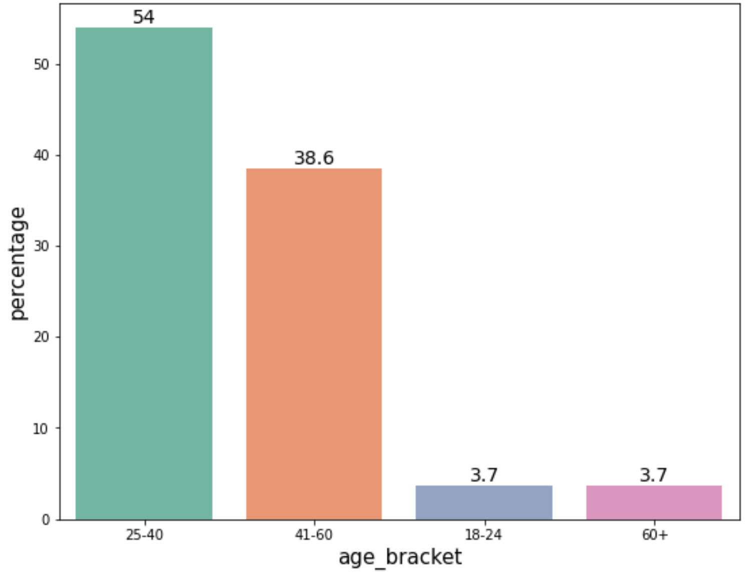
marital	%
married	59.4
single	29.4
divorced	11.1
unknown	0.2

job	%
white-collar	34.1
blue-collar	19.2
technician	16.0
other	12.6
pink-collar	11.4
self-employed	3.8
entrepreneur	2.9

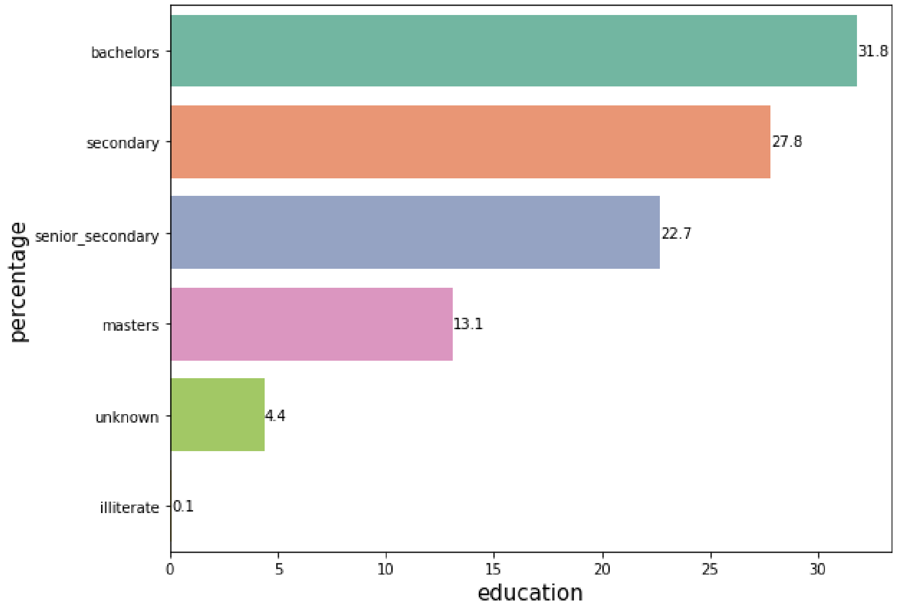
has_housing_loan	%
yes	52.9
no	44.8
unknown	2.3

has_personal_loan	%
no	83.4
yes	14.3
unknown	2.3

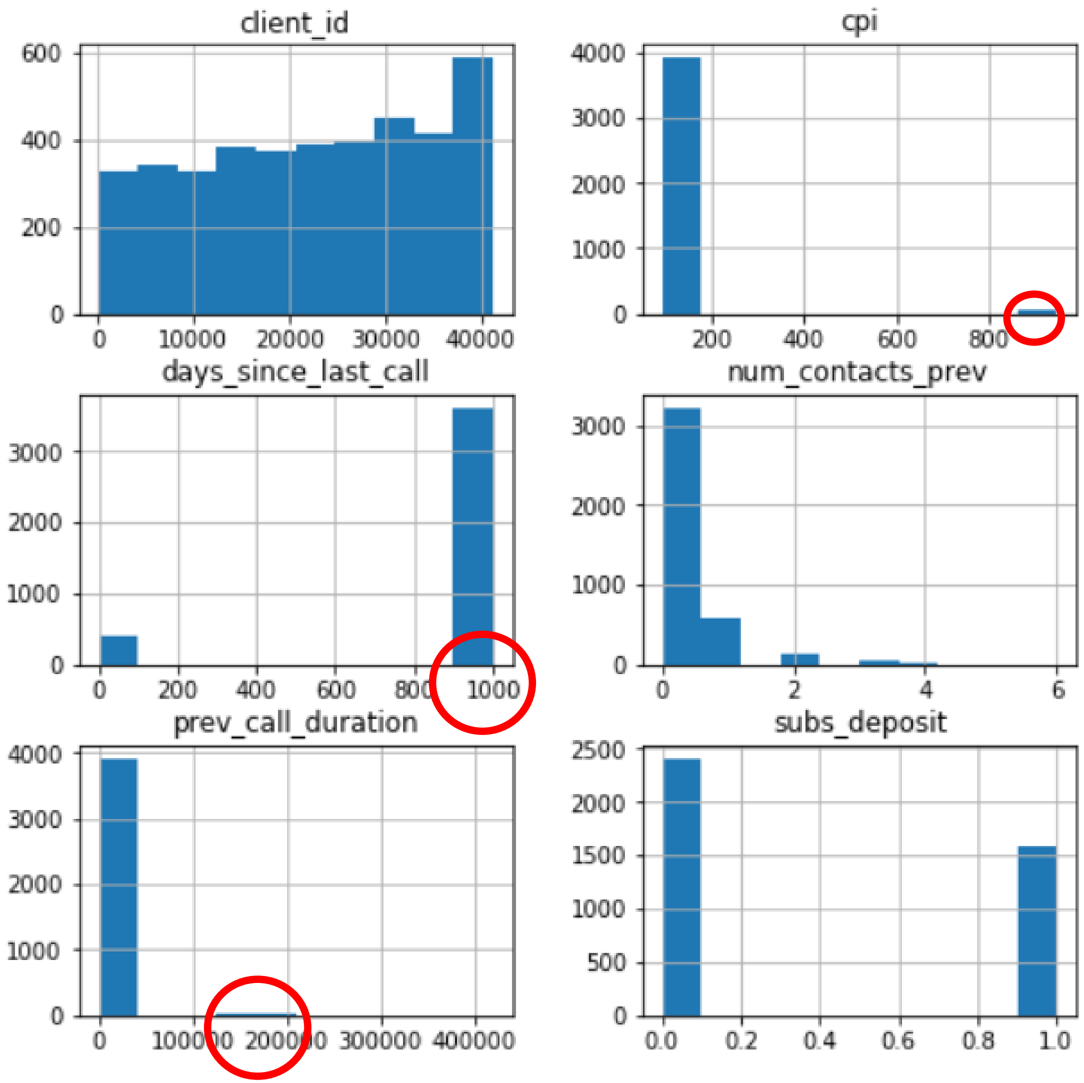
age bracket



education



Exploratory Data Analysis



Notes on selected information

- missing values for days since last call were coded 999
- some unusually high cpi values
- anomalous prev call duration values

Information on previous customer interactions suggest that:

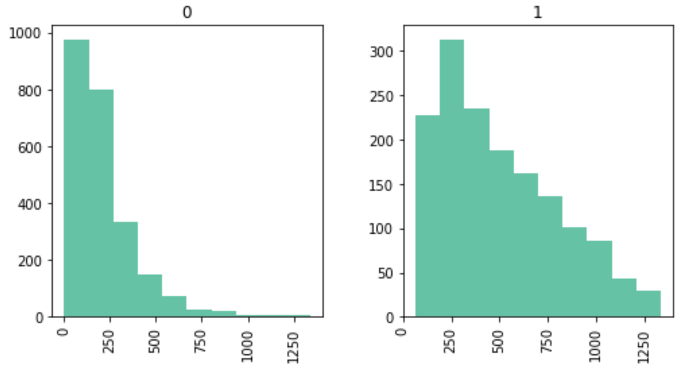
- median previous call duration is ~4 minutes (231 seconds)
- mean previous contacts is zero
- median days since last call was 6 days (removing values coded 999)

Exploratory Data Analysis – Statistical Tests

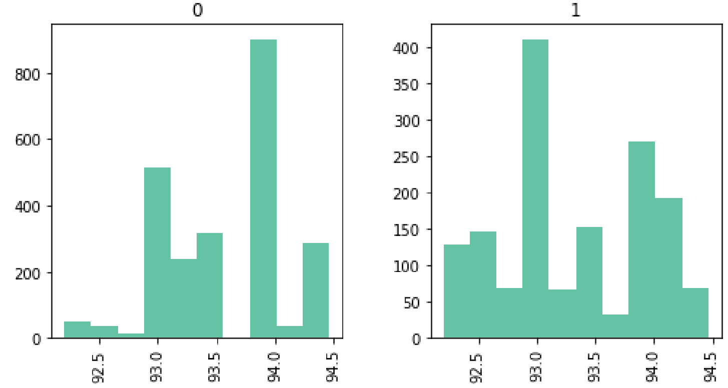
all continuous vars ('prev_call_duration', 'days_since_last_call', 'num_contacts_prev', 'cpi') were statistically different on a t-test grouped by the outcome var (subs_deposit) *

*^ after data cleaning client_id, cpi and prev_call_duration were statistically significant at p value of 0.05

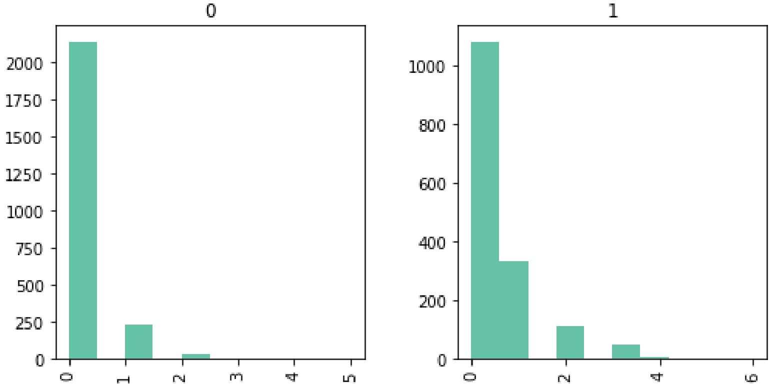
prev_call_duration



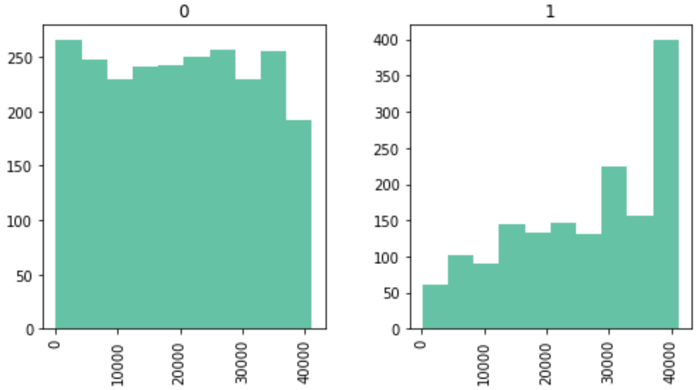
cpi



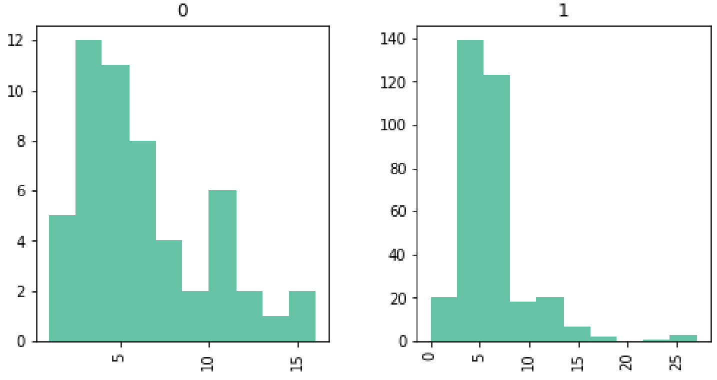
num_contacts_prev



client_id



days_since_last_call



client_id mistakenly omitted from initial analyses

Exploratory Data Analysis – Statistical Tests

chi-square test indicates that there is statistical relationship b/w sub_deposits and poutcome

Clients who chose not to subscribe to a deposit were much more likely to have a greater representation of 'nonexistent' for previous outcome 'poutcome' than those that did subscribe (subs_deposit=1)

Chi-square tests for categorical variables

variable	p value
age_bracket	0.995
job	1.000
marital	0.999
education	1.000
has_housing_loan	0.999
has_personal_loan	1.000
poutcome	0.915
contact_date	1.000

poutcome	%
nonexistent	80.5
failure	10.5
success	9.1

Feature Engineering – anomalous data for days_since_last_call, cpi, prev_call_duration

	client_id	age_bracket	job	marital	education	has_housing_loan	has_personal_loan	prev_call_duration	days_since_last_call	num_contacts_prev	poutcome	contact_date	cpi	subs_deposit
0	41020	41-60	white-collar	divorced	bachelors	yes	no	283	3	1	success	07/09/18	92.379	1
1	23720	60+	other	divorced	secondary	no	yes	169	6	2	success	05/07/18	94.215	1
2	29378	41-60	white-collar	married	bachelors	no	no	552	999	0	nonexistent	01/08/18	93.444	1
3	36636	25-40	technician	single	senior_secondary	yes	yes	206	999	0	nonexistent	02/11/18	93.200	0
4	38229	18-24	white-collar	single	bachelors	no	no	341	999	0	nonexistent	04/04/18	93.075	1

- Binary indicator variables were created for each categorical variable

- Created two outlier (anomalous data) features, instead of replacing with missing values:
    - cpi_outlier (cpi > 800)
    - prev_call_outlier (prev_call_duration > 100,000)

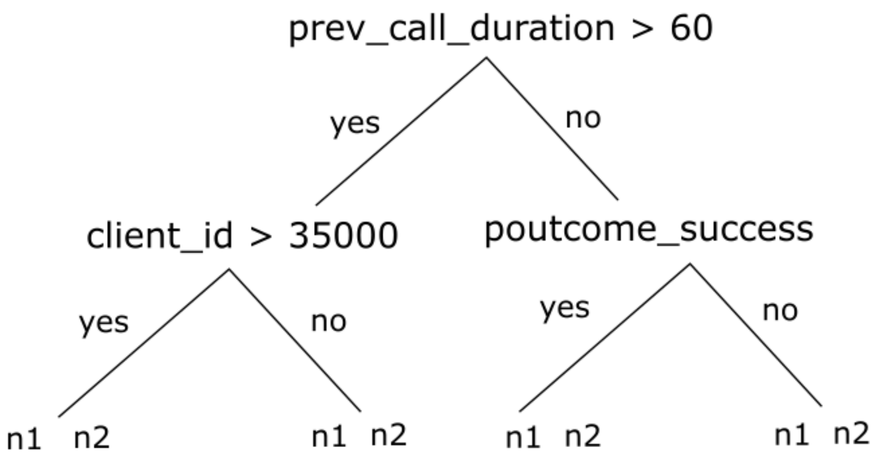
- drop contact_date as a timestamp variable because information is reflected in days_since_last_call

- for days_since_last_call (999), a new 'no contact' feature is not necessary given the 'num_contacts_prev' features already captures when the customer has not had previous contact

Machine Learning Model

- The problem is essentially a binary classification model. i.e. trying to predict whether the outcome is a zero or one, based on a number of predictors

- Decision tree family of models tend to perform well on this type of problem **and are interpretable**



- Random forests sample with replacement to overcome overfitting & have random subset of features

List of features

client_id, cpi, days_since_last_call, num_contacts_prev, prev_call_duration

feature 0: client_id',
'feature 1: cpi',
'feature 2: days_since_last_call',
'feature 3: num_contacts_prev',
'feature 4: prev_call_duration',
'feature 5: age_bracket_18-24',

age_bracket

'feature 6: age_bracket_25-40',
'feature 7: age_bracket_41-60',
'feature 8: age_bracket_60+',
'feature 9: education_bachelors',

education

'feature 10: education_illiterate',
'feature 11: education_masters',
'feature 12: education_secondary',
'feature 13: education_senior_secondary',
'feature 14: education_unknown',

has_housing_loan

'feature 15: has_housing_loan_no',
'feature 16: has_housing_loan_unknown',
'feature 17: has_housing_loan_yes',

has_personal_loan

'feature 18: has_personal_loan_no',
'feature 19: has_personal_loan_unknown',
'feature 20: has_personal_loan_yes',

type of job

'feature 21: job_blue-collar',
'feature 22: job_entrepreneur',
'feature 23: job_other',
'feature 24: job_pink-collar',
'feature 25: job_self-employed',
'feature 26: job_technician',
'feature 27: job_white-collar',

marital status

'feature 28: marital_divorced',
'feature 29: marital_married',
'feature 30: marital_single',
'feature 31: marital_unknown',

previous outcome

'feature 32: poutcome_failure',
'feature 33: poutcome_nonexistent',
'feature 34: poutcome_success',

outlier features

'feature 35: cpi_outlier',
'feature 36: prev_call_outlier'

Machine Learning Model

	Prediction Success	Prediction Fail
Actual Success	TP	FN
Actual Fail	FP	TN

Reference: dataiku ml classification ppt

precision : true positive (TP) / (True Positive + False Positive) percentage of TP correct from all guesses

recall: True Positive / (True Positive + False Negative) percentage of TP correct from all positives

f1 is the harmonic mean of precision and recall, seeking to strike a balance 2 x (P * R) / (P + R)

hyperparameter tuning

- looped over the number of features to include in the decision tree per the random forest method: 10 to 35 in increments of 5
- looped over the depth of the decision tree to stop running at: 5 to 25 in increments of 5
- defaults for other parameters, number of samples (random forests) = 4,000, gini impurity = 1 – p1^2 – p2^2

ML Model Results

With a 20% validation dataset, the best two scoring models with the outlier features were:

n_est: 35, md: 15, precision: 0.82, recall: 0.82, f1 0.819
n_est: 25, md: 10, precision: 0.82, recall: 0.81, f1 0.815

preferred model using 25 features and max depth of 10. slight concern with 35 features and 15 max depth of overfitting

^caveat that gradient boosting models such as xgboost would probably provide a better overall result

accuracy

y	0.0	1.0
yhat		
0.0	53.125	7.00
1.0	7.625	32.25

accuracy (TP + TN) / (P + N)of 85.375

Feature Importance

Top 20 most important features

feature	mean decrease in impurity
prev_call_duration	0.397
client_id	0.129
cpi	0.120
poutcome_success	0.063
days_since_last_call	0.035
num_contacts_prev	0.031
prev_call_outlier	0.019
poutcome_nonexistent	0.018
age_bracket_60+	0.016
job_other	0.014
marital_single	0.011
poutcome_failure	0.010
education_secondary	0.009
education_bachelors	0.009
age_bracket_41-60	0.009
age_bracket_18-24	0.009
has_housing_loan_yes	0.008
marital_married	0.008
age_bracket_25-40	0.008
has_housing_loan_no	0.007

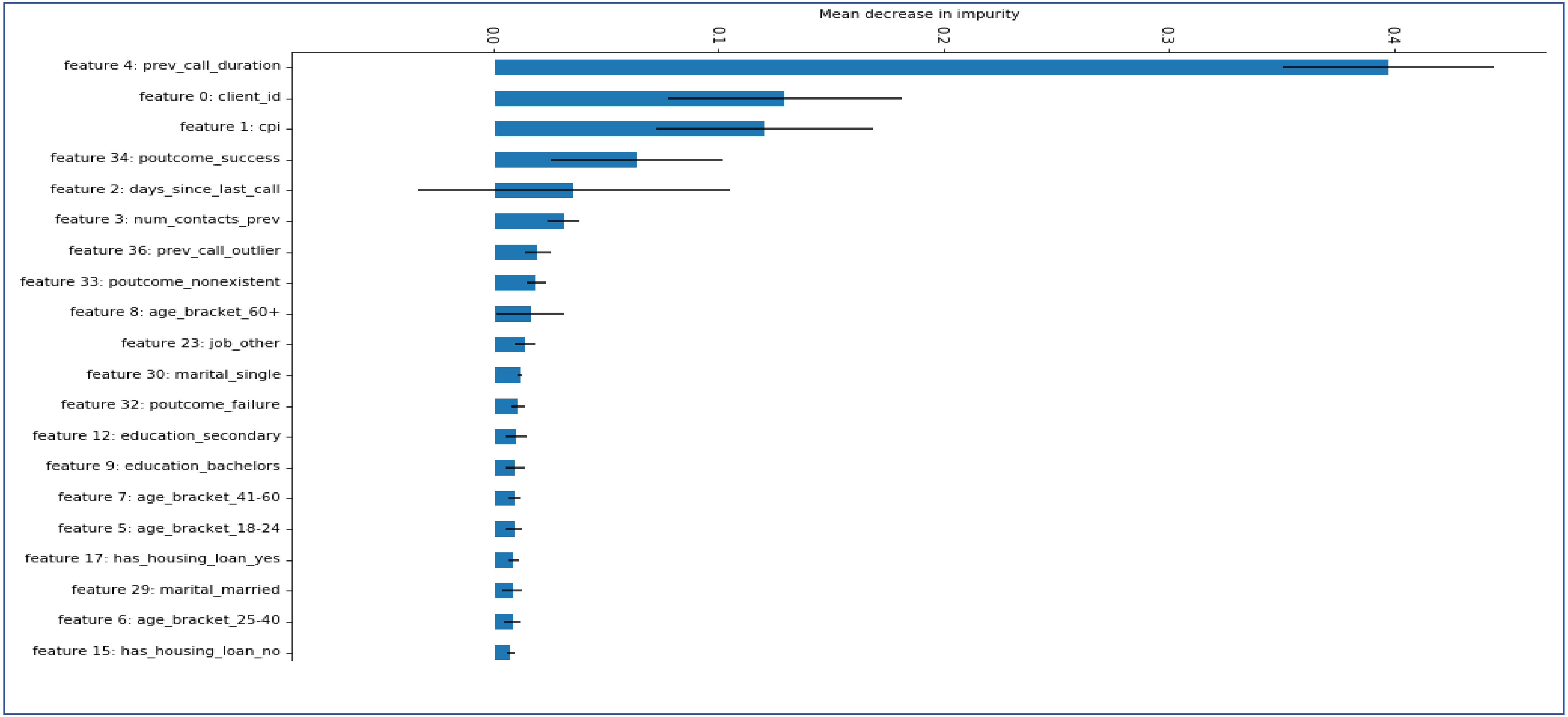
- prev call duration is the most important, with more than double the magnitude of the next feature, at 0.397. This is the amount of time a customer has stayed on the last call

- 2nd is client_id at 0.129, which may reflect how recently the customer joined the bank if the client_id is related to time

- next 3 most important features are: cpi (0.120), poutcome_success (0.063) and days_since_last_call (0.035), ranging from: 0.12 to 0.035

- 6th most important feature were the number of previous contacts with a Mean Decrease in Impurity of 0.306. The remaining features have magnitudes less than 0.02 decrease in impurity (gini)

Feature Importance



Summary

Sample Submission

For the sample submission of 1,000 clients, the model identified 425 clients who were likely to subscribe to the new term deposit product, and 575 who were not

Recommendation

Analysis of provided data indicates that Lending Bank's sales manager should target:

- customers with longer prev_call_duration (222.3 vs 2901.6)
- recent customers (from client_id 20085.9 vs 25984.7)
- greater success with lower cpi (93.3 vs 93.6)
- customers with success at previous outcome
- customers with lower days_since_last_call (5.7 vs 6.0)
- higher num_contacts_prev (0.5 vs 0.1)

Appendix - Exploratory Data Analysis

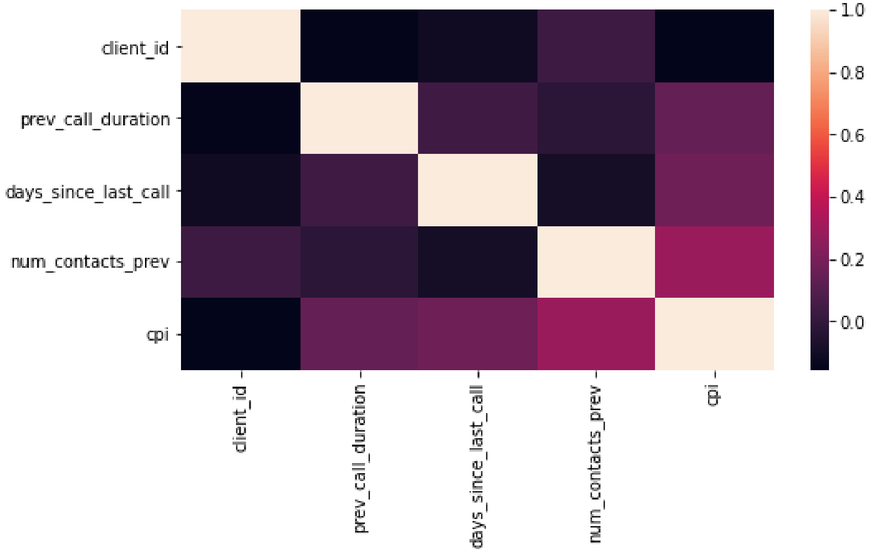
Expectations from **EDA**, variables with the strongest relationship to subscribing to a term deposit were:

- prev_call_duration
- days_since_last_call^
- num_contacts_prev^
- cpi^
- poutcome

	count	mean	std	min	5%	50%	95%	97.5%	99%	max
cpi	4000.0	107.3	107.9	92.2	92.4	93.4	94.5	94.5	946.0	947.7

	count	mean	std	min	5%	50%	95%	97.5%	99%	max
client_id	4000	22431	12053	17	2544	23336	39701	40633	40922	41186
prev_call_duration	4000	3871	26081	2	50	237	1030	1269	157400	419900
days_since_last_call	4000	903	293	0	6	999	999	999	999	999
num_contacts_prev	4000	0	1	0	0	0	2	2	3	6

days_since_last_call	386.0	5.7	3.8	0.0	2.0	6.0	13.0	15.0	18.6	27.0



	client_id	prev_call_duration	days_since_last_call	num_contacts_prev	cpi
client_id	1.0	-0.1	-0.1	0.0	-0.2
prev_call_duration	-0.1	1.0	0.0	-0.0	0.1
days_since_last_call	-0.1	0.0	1.0	-0.1	0.2
num_contacts_prev	0.0	-0.0	-0.1	1.0	0.3
cpi	-0.2	0.1	0.2	0.3	1.0