PROBLEM STATEMENT

To predict if a particular customer will subscribe to a term deposit in a bank.

Relevance

- Increase in marketing strategies by various banks in recent times has reduced the effect on customers.
- Stiff competition from national and international players vying for customers; hence resource allocation optimization is required.
- According to recent study[1], less than 1% of the contacts subscribe to a term deposit.

Data introduction

- This is a modified version of the marketing dataset posted on UCI repository by a Portuguese bank.
- The aim of the project is related to a direct marketing campaign where the goal is to predict if a client will subscribe a term deposit or not given a set of relevant information regarding the contact.

[1] Who will subscribe a term deposit, J. Chen et. al. Advanced Data Analysis, Department of Statistics, Columbia University,

Who will benefit?

- Various banks and specifically their marketing groups will likely benefit from this study.

BANK OF AMERICA

citibank

WELLS FARGO

CHASE

Deutsche Bank

What factors likely contribute to customer subscription?

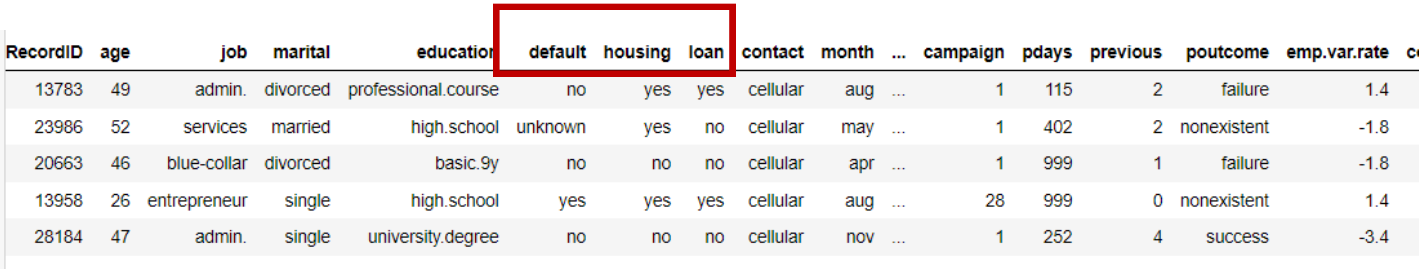
Factors expected to influence model

- Employment
- Credit history
- Loan history
- Outcome of previous marketing campaign
- Last contact duration
- Education

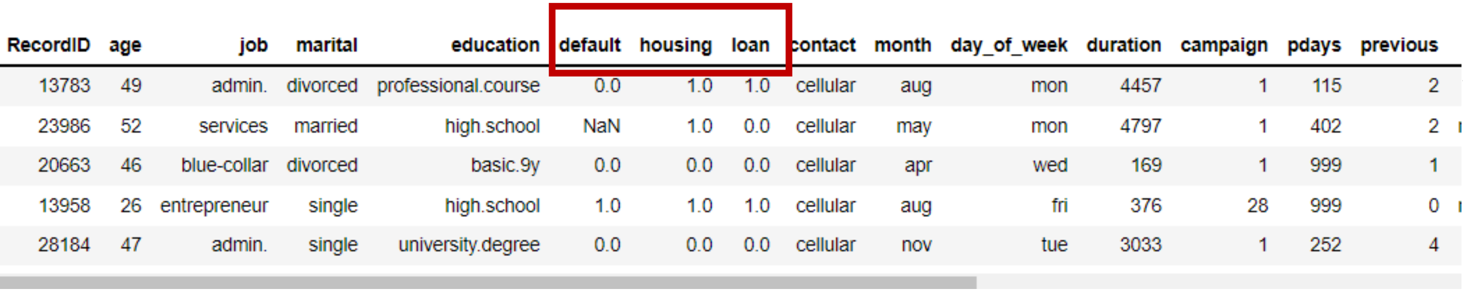
Factors expected to have less/no effect

- Day and month of last contact
- Employment variation rate
- Consumer Price Index
- Euribor rate (quarterly)
- Consumer price index
- Outcome of last contact

Data Wrangling



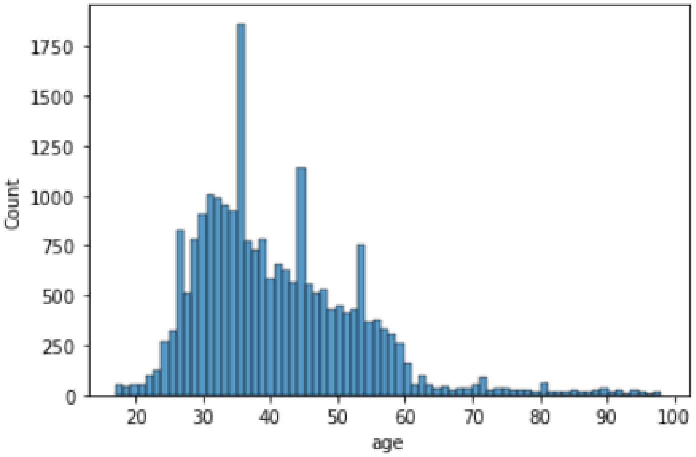
RecordID	age	job	marital	education	default	housing	loan	contact	month	...	campaign	pdays	previous	poutcome	emp.var.rate	c
13783	49	admin.	divorced	professional.course	no	yes	yes	cellular	aug	...	1	115	2	failure	1.4	
23986	52	services	married	high.school	unknown	yes	no	cellular	may	...	1	402	2	nonexistent	-1.8	
20663	46	blue-collar	divorced	basic.9y	no	no	no	cellular	apr	...	1	999	1	failure	-1.8	
13958	26	entrepreneur	single	high.school	yes	yes	yes	cellular	aug	...	28	999	0	nonexistent	1.4	
28184	47	admin.	single	university.degree	no	no	no	cellular	nov	...	1	252	4	success	-3.4	

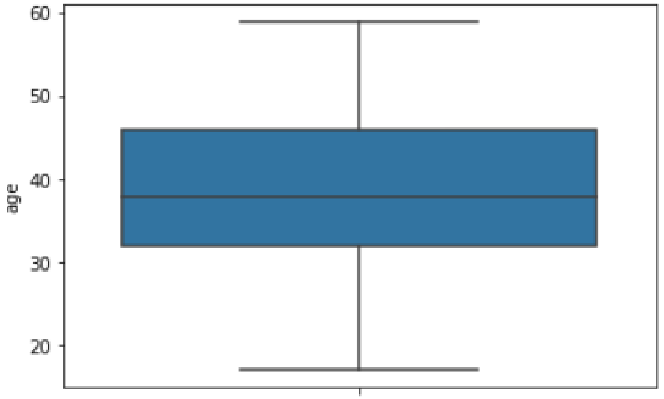


RecordID	age	job	marital	education	default	housing	loan	contact	month	day_of_week	duration	campaign	pdays	previous
13783	49	admin.	divorced	professional.course	0.0	1.0	1.0	cellular	aug	mon	4457	1	115	2
23986	52	services	married	high.school	NaN	1.0	0.0	cellular	may	mon	4797	1	402	2
20663	46	blue-collar	divorced	basic.9y	0.0	0.0	0.0	cellular	apr	wed	169	1	999	1
13958	26	entrepreneur	single	high.school	1.0	1.0	1.0	cellular	aug	fri	376	28	999	0
28184	47	admin.	single	university.degree	0.0	0.0	0.0	cellular	nov	tue	3033	1	252	4

Steps:

1. *Data was checked for null and duplicate values.*
2. *Features with yes/no values were converted to 1/0.*
3. *The resulting NaN values were dropped since these featured were deemed to be important and cannot be imputed.*
4. *Correlation matrix showed weak correlation with ‘default’, ‘campaign’, ‘emp.var.rate’, ‘euribor3m’, and ‘nr.employed’.*
5. *Age distribution was found to be Gaussian and hence outliers were removed.*





Data Pre-processing

- *Since, distribution of numeric features were not Gaussian, MinMaxScaler was used to standardize the features instead of StandardScaler.*
- *The module “get_dummies” function was called to OneHotEncode all categorical features.*
- *After pre-processing, there were total of 16100 rows X 57 columns.*

Initial Models

- *Initial models were built with logistic regression, decision trees and random forest. The corresponding accuracies have been listed below:*

Models	Accuracy
Logistic Regression	0.8759
Decision Trees	0.856
Random Forest	0.892

- *Based on accuracy of initial models, further optimization was carried out on Logistic Regression, Decision Trees Random Forest models. A Gradient Boost model was also investigated.*

Modelling

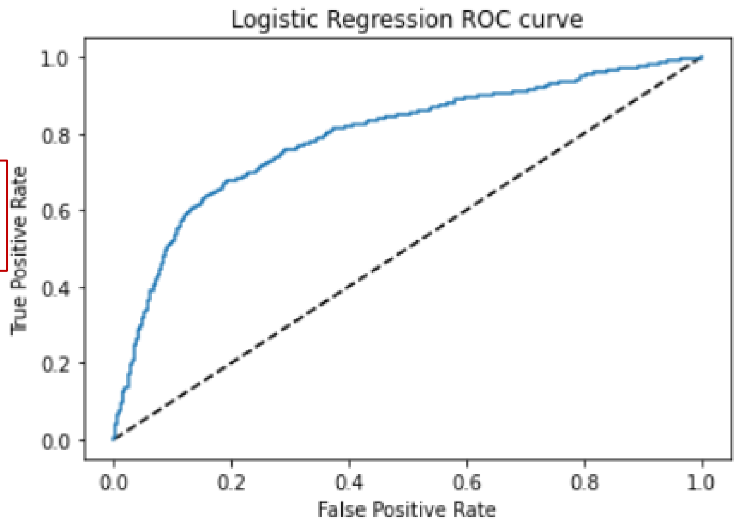
- Based on initial models, further Hyperparameter Tuning, Grid Search and Cross Validation was carried out using Logistic Regression, Random Forest, Decision Trees and Gradient Boost.

Model	Optimized Hyperparameter	Accuracy
Logistic Regression	C= 0.1	0.88
Random Forest Classifier	'n_estimators': 275, 'min_samples_split': 2, 'min_samples_leaf': 4, 'max_features': 'auto', 'max_depth': 10, 'bootstrap': True	0.892
Decision Tree Classifier	'criterion': 'entropy', 'max_depth': 10, 'min_samples_leaf': 3, 'min_samples_split': 9	0.886
Gradient Boost Classifier	'learning_rate': 0.1, 'max_depth': 5, 'n_estimators': 50	0.897

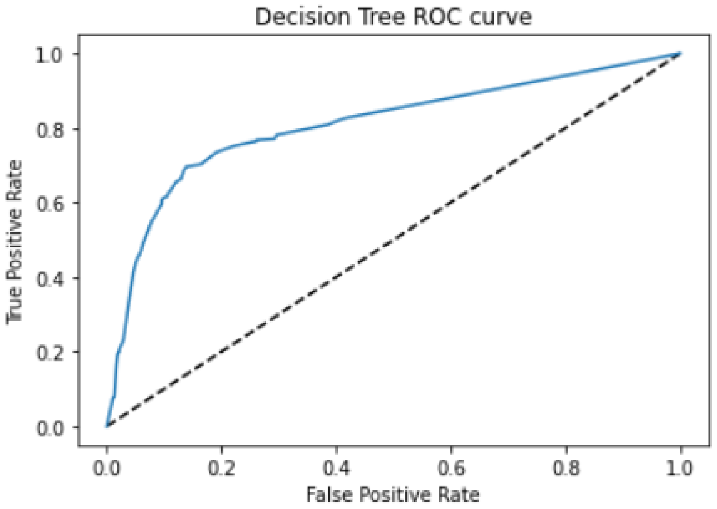
- After hyperparameter tuning, the Gradient Boost Classifier gave the highest accuracy of 0.897.

Modelling

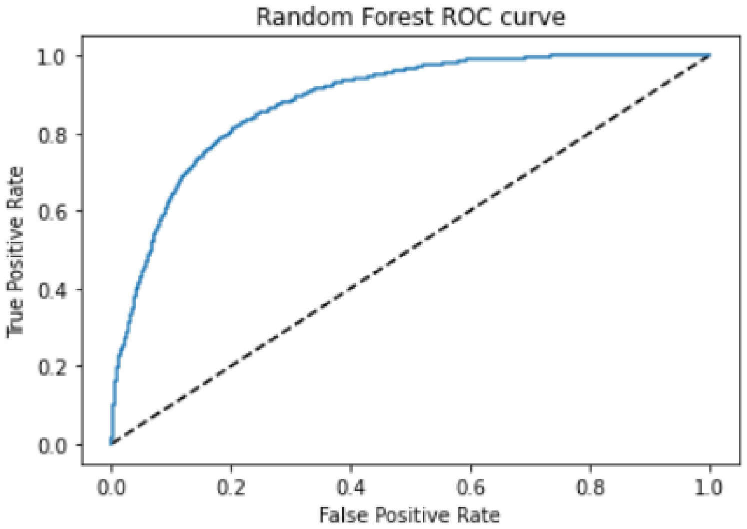
- The ROC and AUC was calculated for each of the models.



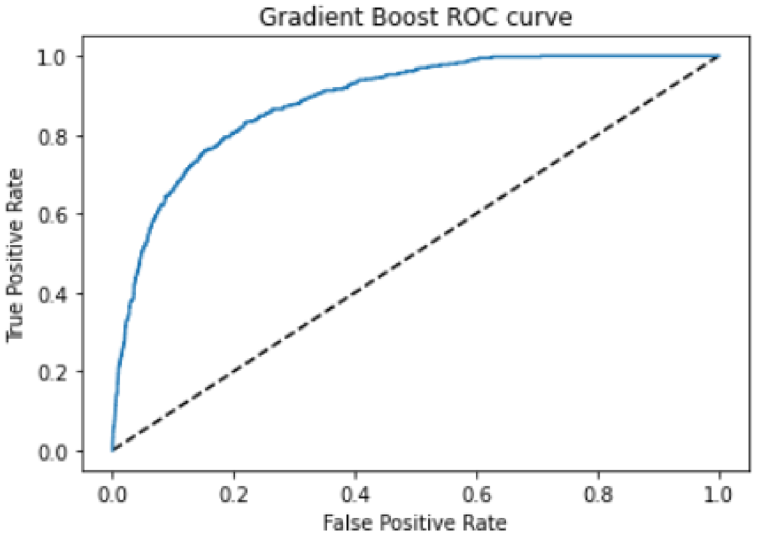
Logistic Regression AUC 0.791



Decision Tree AUC 0.808



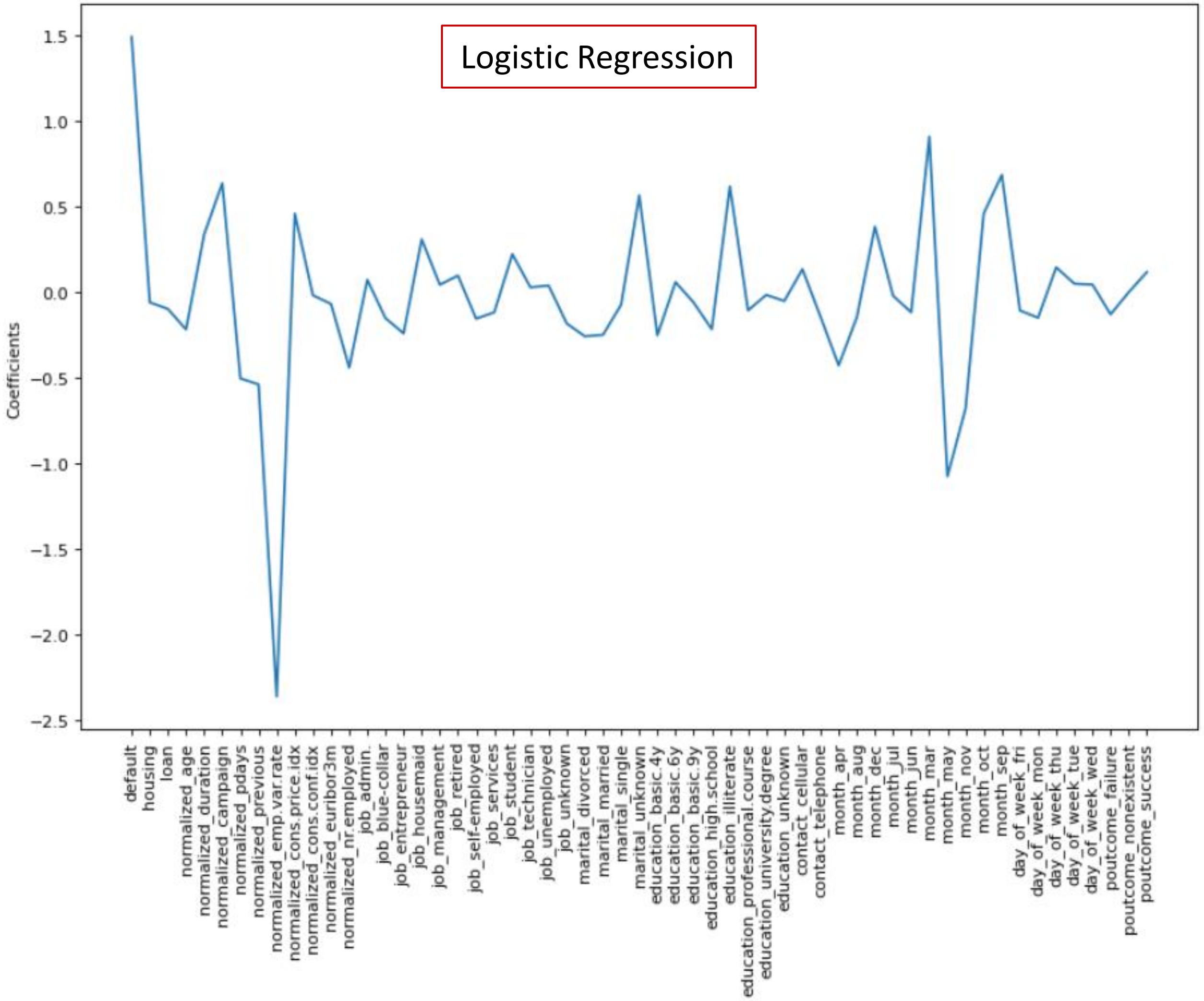
Random Forest AUC 0.883



Gradient Boost AUC 0.888

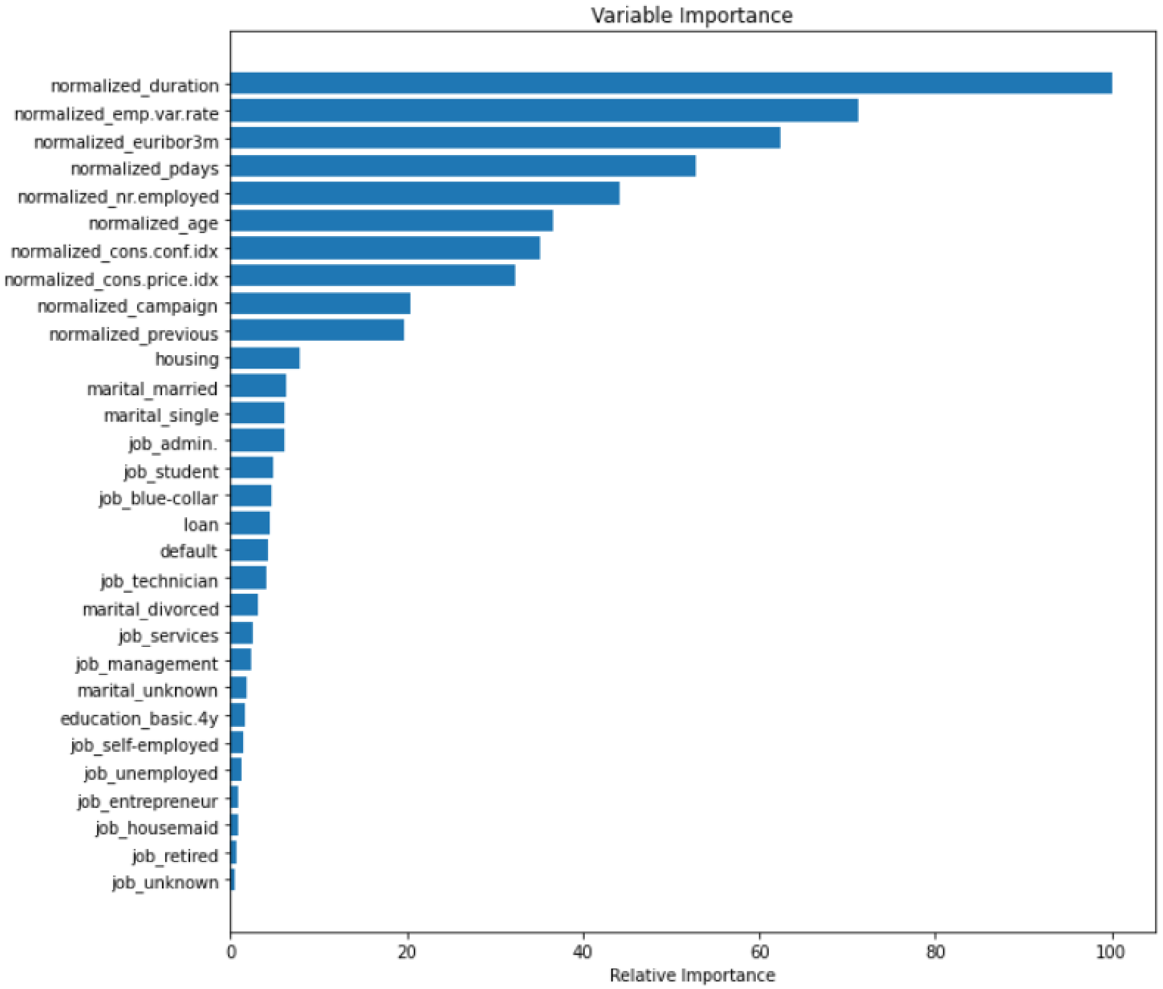
Feature Importance

- *The feature importance plots were compiled for Random Forest and Gradient Boost model. For comparison, the Logistic Regression coefficients were also plotted*



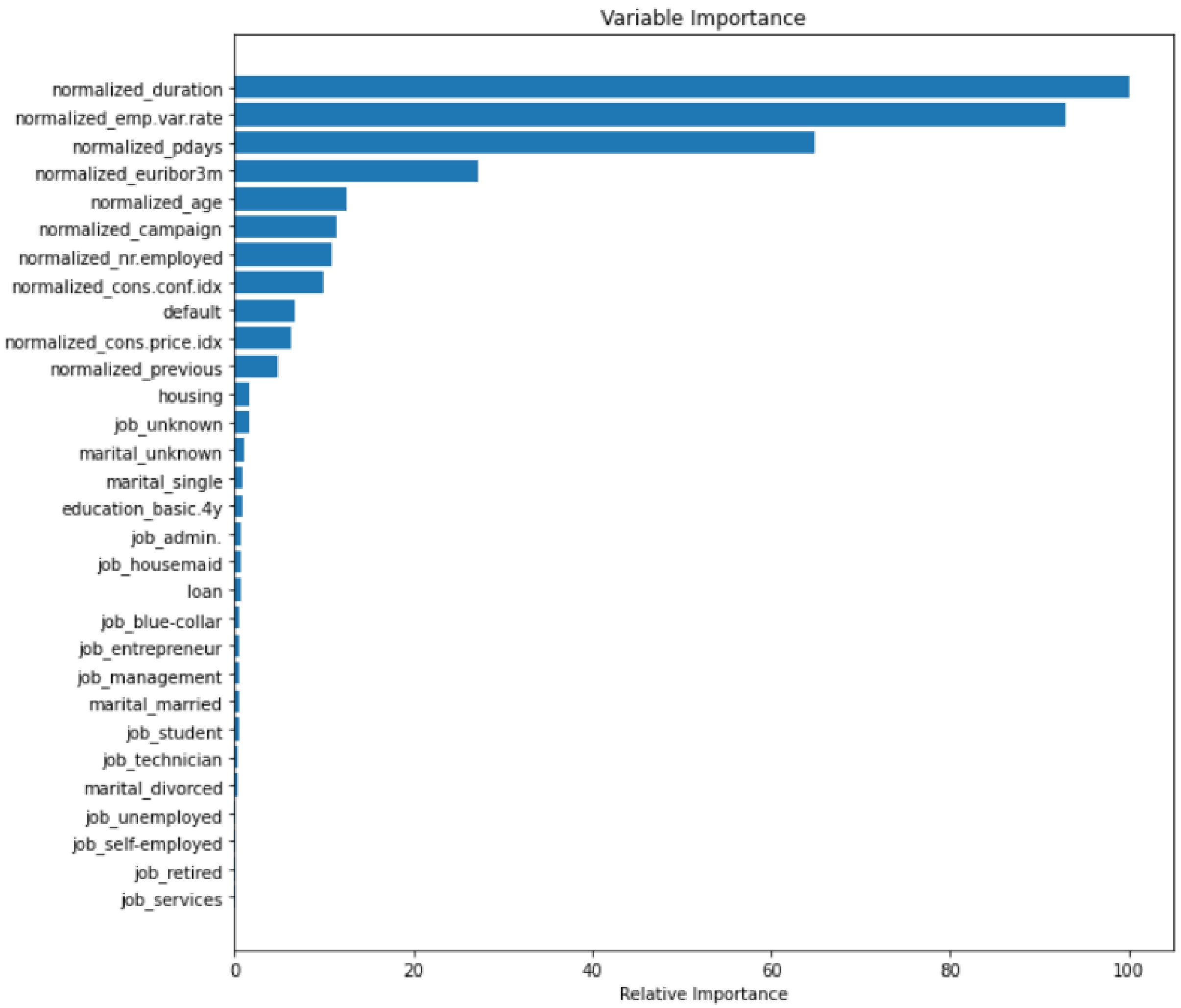
Logistic Regression

Random Forest



Feature Importance

Gradient Boost



- *Top 5 relevant features for Random Forest model and Gradient Boost model in descending order.*

Random Forest

- *Duration: last contact duration*
- *Employment variation rate*
- *Euribor3m*
- *Pdays: No of days after previous contact*
- *NR employed: No. of employees*

Gradient Boost

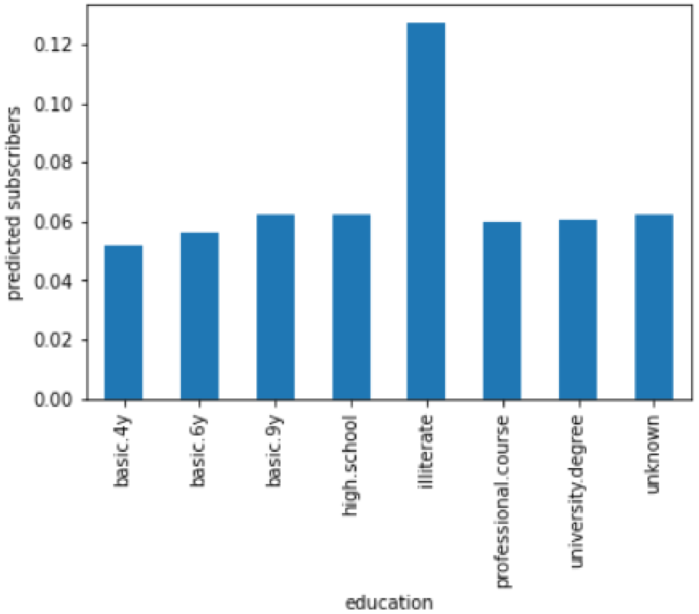
- *Duration: last contact duration*
- *Employment variation rate*
- *Pdays: No of days after previous contact*
- *Euribor3m*
- *Age*

- As seen, four out of five parameters are common for both models.

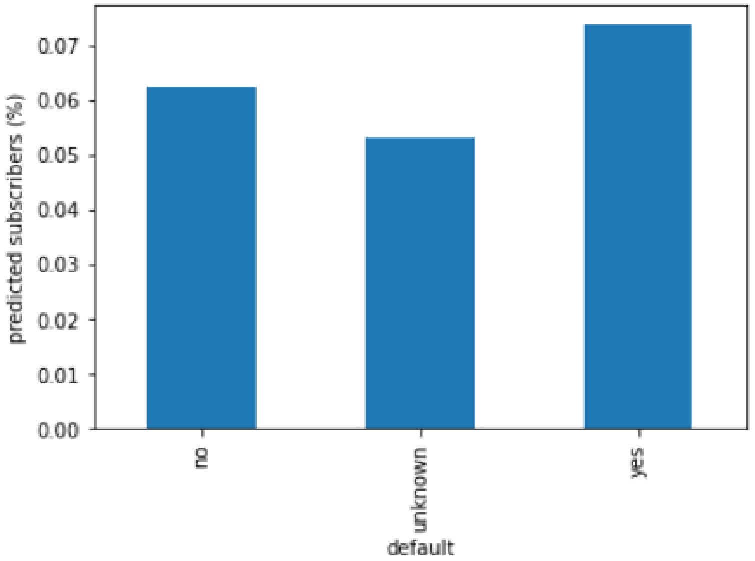
Investigating Test Data

- *The optimized Gradient Boost model was run on test data.*
- *The model predicts 456 contacts out of total of 5626 should subscribe for term deposit.*
- *This results is 8.1% subscription rate which is much larger than traditional ~1% stated in the beginning.*

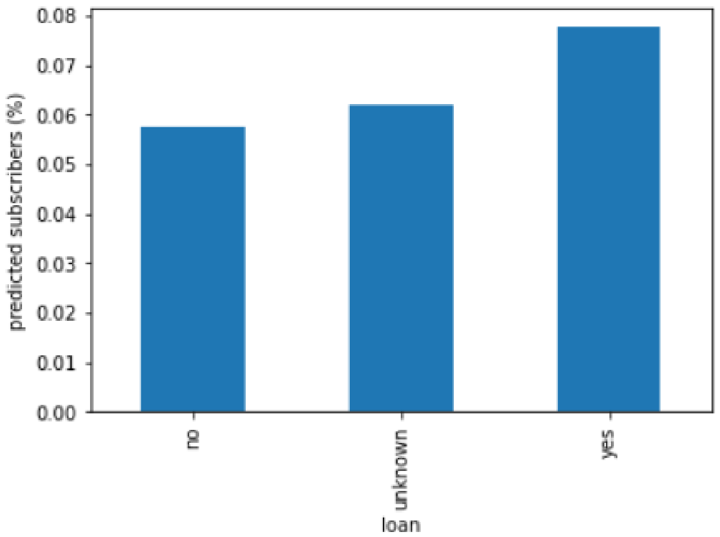
- *Several features were investigated to understand if the factors that were thought to influence predictions are relevant or not.*



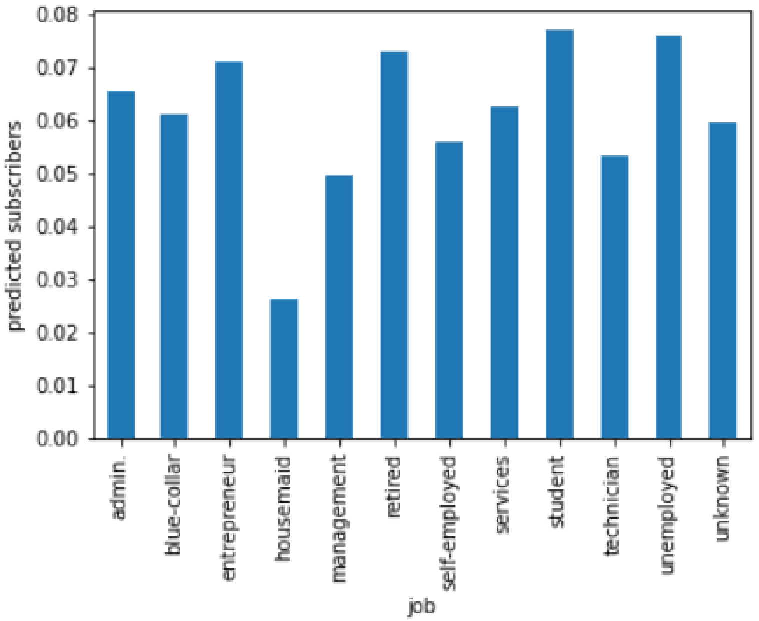
Though education is not a relevant feature, it seems contacts who are listed as "illiterate" subscribe higher than other education group from a percent standpoint



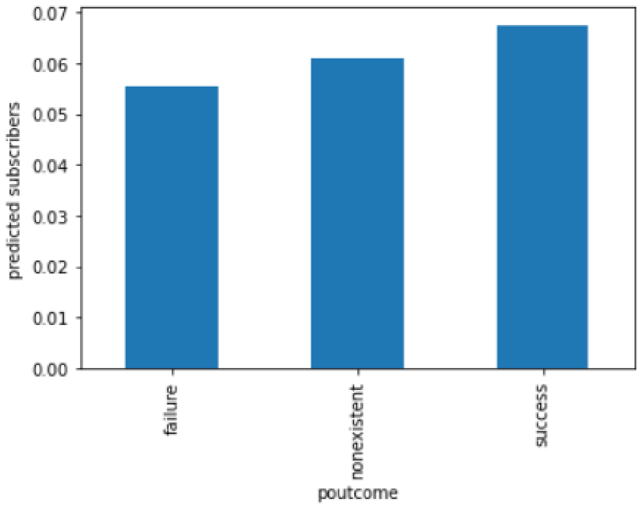
Contacts with default credit history subscribe to term deposits in the same percentage has contacts with no defaults



Contacts with previous loan in account, subscribe in higher numbers than contacts with no loan.



Except for housemaids, contacts from other jobs seem to subscribe in similar ratios



Previous marketing campaign outcome is irrelevant when it comes to new subscriptions.