deti universidade de aveiro
departamento de eletrónica, telecomunicações e informática

Bank Marketing

Dataset

Assignment 1

Machine Learning Fundamentals

Miguel Cruzeiro (107660)

Miguel Figueiredo (108287)

Professor

Petia Georgieva

Context

- Dataset from a Portuguese Bank's Direct **Marketing Campaign**
- **Binary Classification** Problem
- Predict whether a client will **subscribe to a term deposit**

deposit			
age	job	marital	education
default	balance	housing	loan
contact	day	month	duration
campaign	pdays	previous	poutcome

State of the art

Deteting important features

Using Data Mining Techniques for Detecting the Important Features of the Bank Direct Marketing Data

Tuba Parlar and Songul Kakilli Acaravci

- Reducing the feature set increases the classifier's performance

IEEE

Optimizing bank marketing campaigns

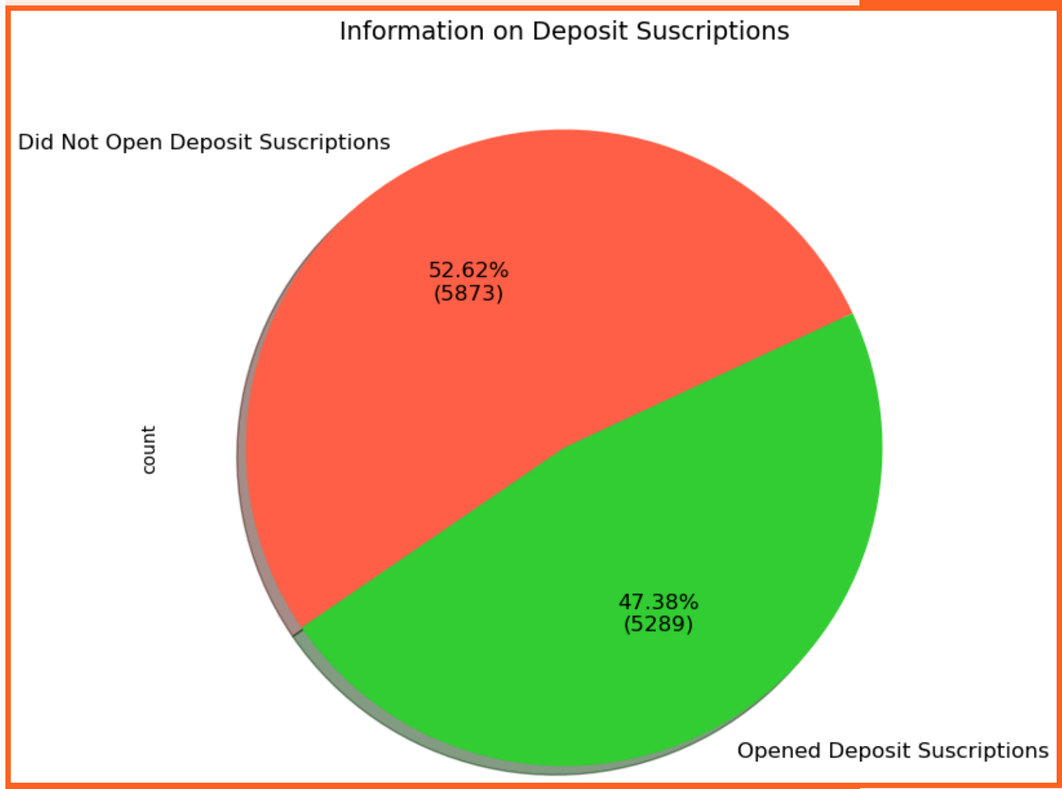
A data-driven approach to predict the success of banktelemarketing

Sergio Moro, Paulo Cortez and Paulo Rita

- Success rate for client subscription is low (12.38%)
- Contacting only 50% of clients ranked by the predictive model, the bank could achieve 79% of successful outcomes.

Dataset Information

- **Fairly Balanced** dataset
- **Accuracy** is a **suitable metric for evaluating the performance** of the model
- **17 Features**
- Information from **11162 individuals**
- **Binary Classification Problem**



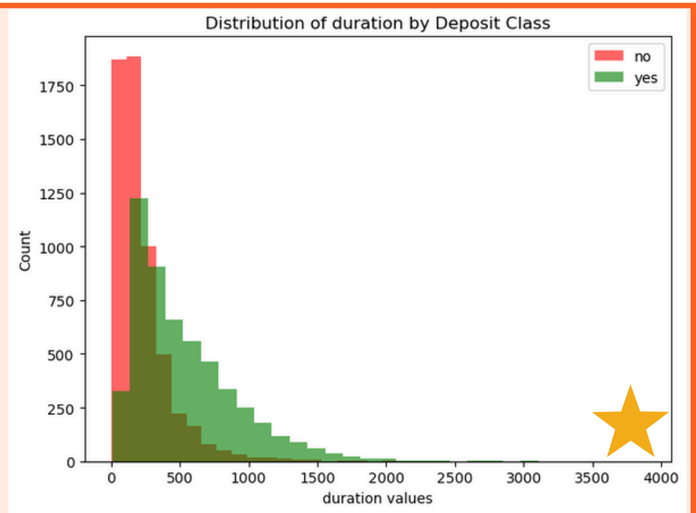
Feature Distribution related to target variable - Num. Features

Slightly **Older Clients**

Higher likelihood to subscribe

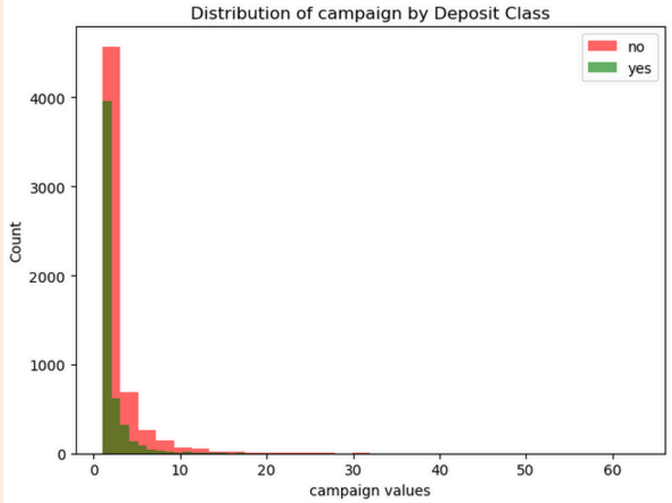
Longer Duration

Higher likelihood to subscribe



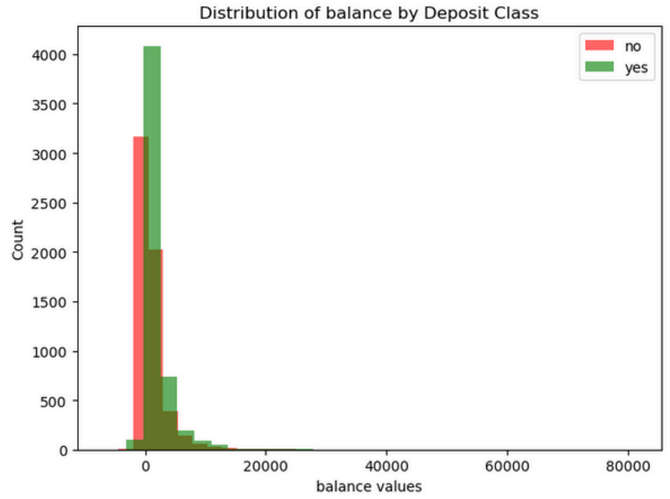
Higher Number of Contacts

Lower likelihood to subscribe



Higher Balance

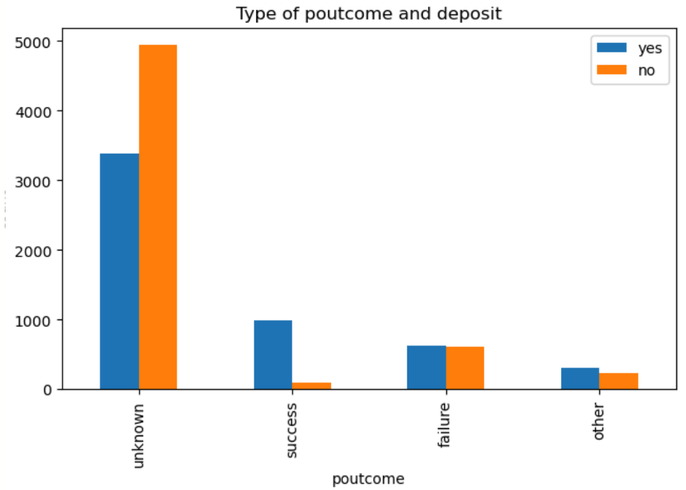
Higher likelihood to subscribe



Feature Distribution related to target variable - Cat. Features

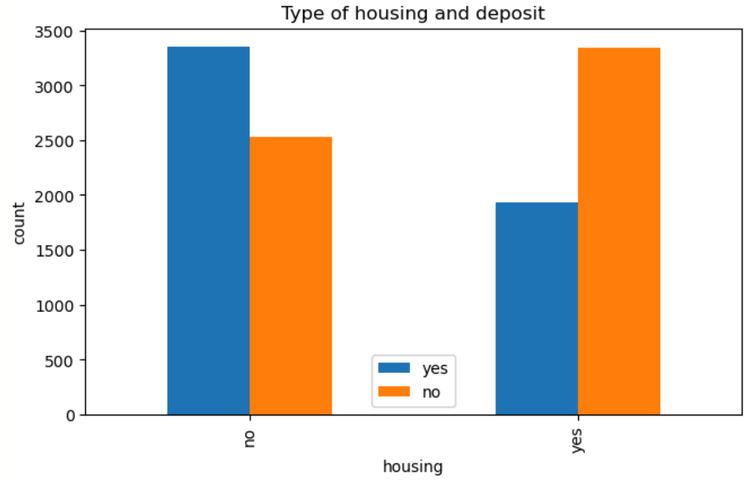
Previously Successful Marketing Campaign

Higher likelihood to subscribe



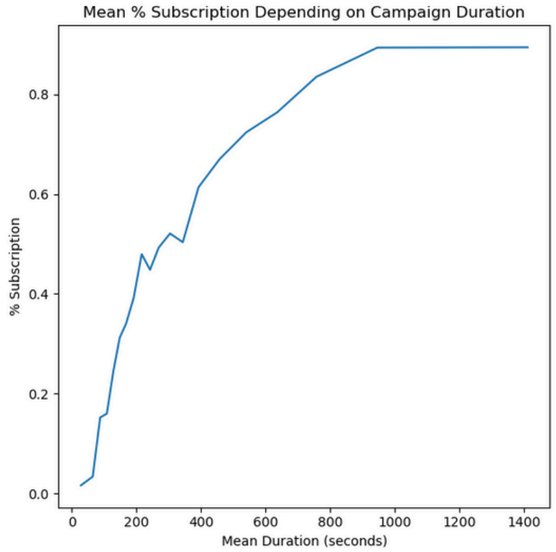
No Housing Loan

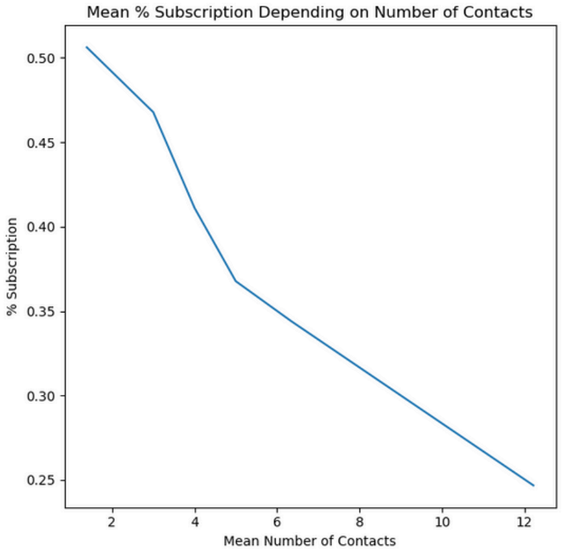
Higher likelihood to subscribe

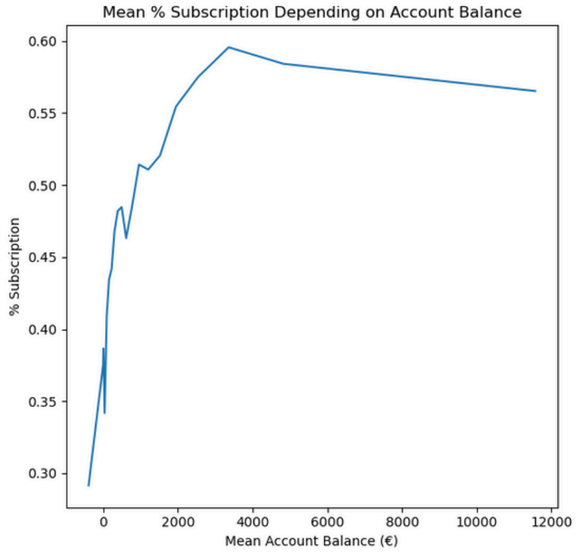


Factors influencing deposit subscriptions

- **Duration of the campaign** and **Account Balance** have a significant impact on the subscription rate
- **Number of contacts** has a negative correlation with subscription rates

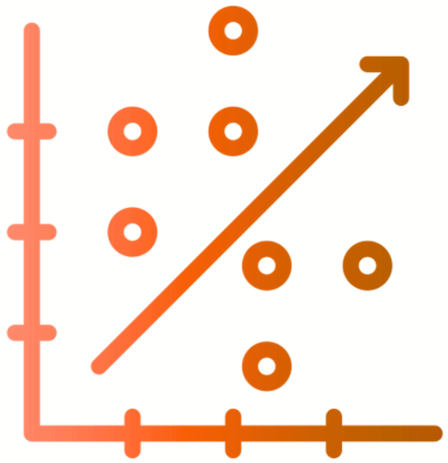


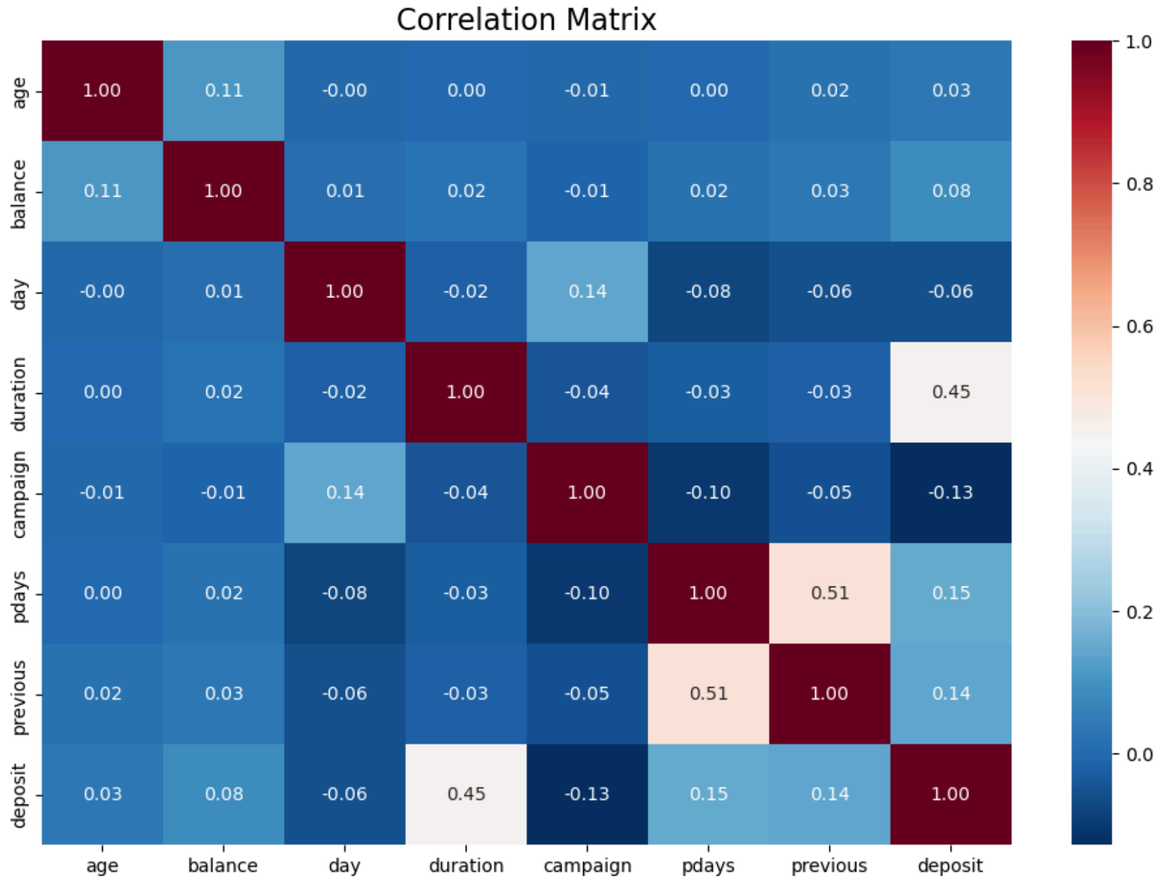




Correlation between features

- **Strong Correlation** Between Call Duration and Deposit Outcome
- **Minimal Correlation** of some variables with Deposit Outcome





Methodology

1

DATA PREPROCESSING

- Handling missing values
- Variable Encoding
- Standardization

2

TRAIN-TEST SPLIT

- Training Set - 80%
- Test Set - 20%

3

MODEL SELECTION

- KNN
- Logistic Regression
- Random Forest
- SVM

4

PARAMETER TUNING

GridSearch on training set to find the best Hyperparameters (scoring = accuracy)

Parameter Tuning

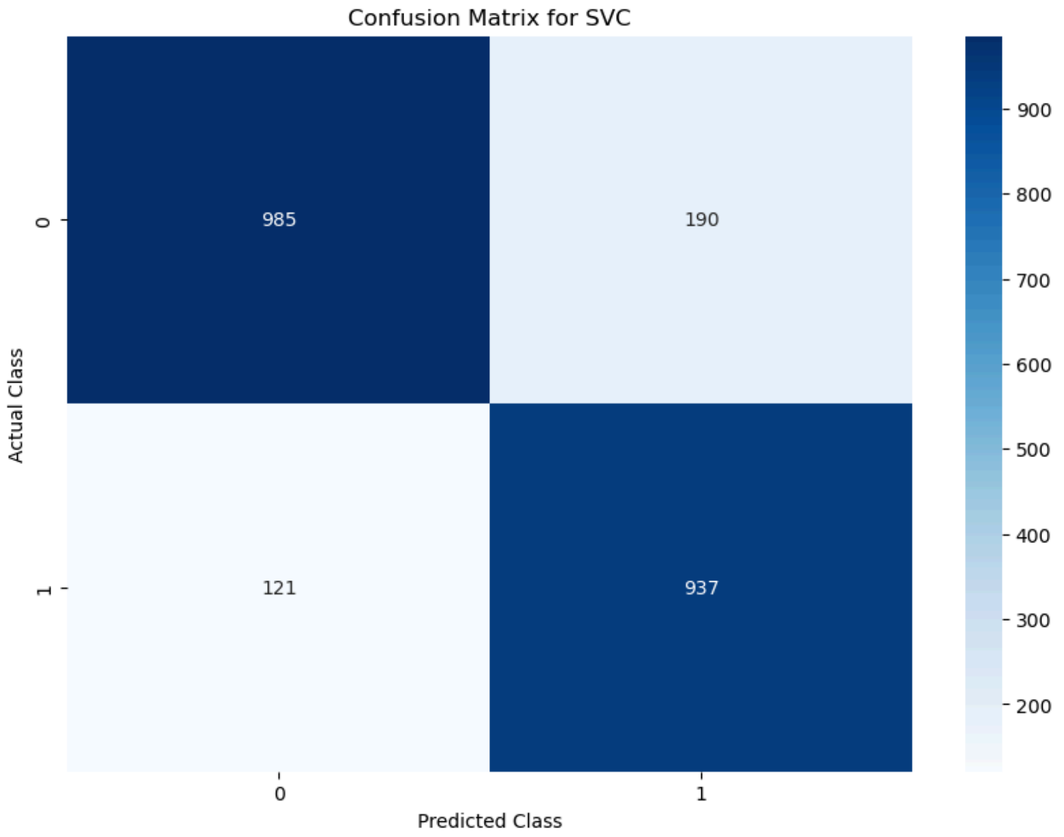
GRID SEARCH ON TRAINING SET

- Exhaustive Search
- Maximizing Accuracy

Model Name	Best Hyperparameters	Best Score (Accuracy)
Knn	**metric**: 'minkowski', **n_neighbors**: 16, **p**: 2, **weights**: 'distance'	0.818
Logistic Regression	**C**: 0.1, **solver**: 'saga', **penalty**: 'l1'	0.828
Random Forest	**max_depth**: 30, **min_samples_leaf**: 2, **n_estimators**: 400	0.855
Support Vector Classifier	**C**: 100, **gamma**: 0.01	0.851

Results

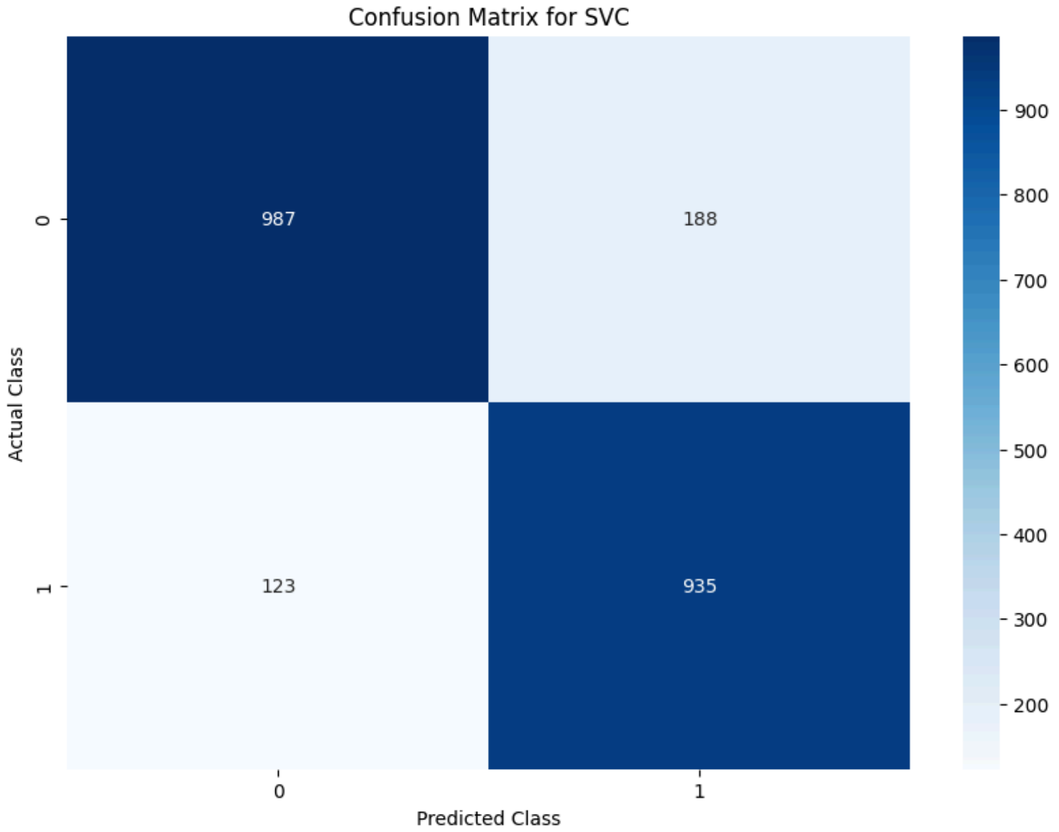
Model Name	Accuracy	Precision	Recall	F1 Score
Knn	0.827	0.827	0.825	0.826
Logistic Regression	0.829	0.828	0.828	0.828
Random Forest	0.858	0.860	0.860	0.858
Support Vector Classifier	0.861	0.861	0.862	0.861



Results after Feature Removal

Model Name	Accuracy	Precision	Recall	F1 Score
Knn	0.830	0.830	0.828	0.829
Logistic Regression	0.830	0.829	0.829	0.829
Random Forest	0.858	0.859	0.860	0.858
Support Vector Classifier	0.861	0.861	0.862	0.861

- **Removed the 8 least Important Features**



Conclusion

- **Best Models:** SVC & Random Forest
- **Parameter Tuning** maximizing accuracy
- **Feature Removal** improved results while:
  - Reducing the **feature scope**
  - Reducing **computational complexity**

Thank you!