EXL

EQ 2022

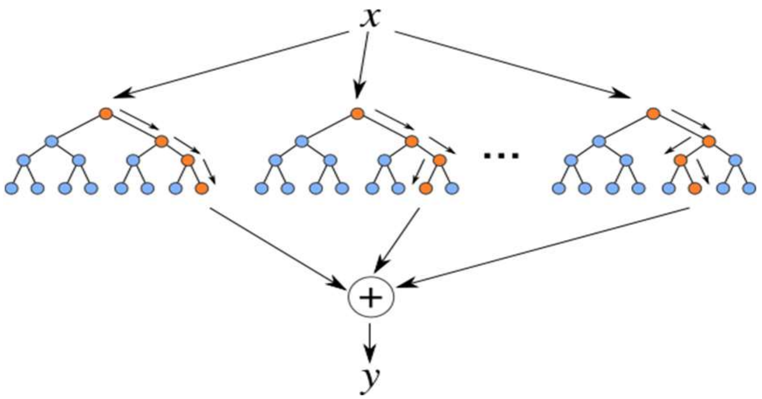
Elite Sapiens (Jadavpur University)

Arnab Dutta
Md Minhaz Rahaman

Understanding the problem

Basic Understanding

- In 2018, ABC's new product 'Term Deposit' was a success and it's subscription benefited the firm.
- So, this year ABC is running a campaign to identify 1000 customers from their database, whom, contacting via telecommunication channels will provide efficient results/ maximum subscription owing to its time and budget constraints.



Random Forest pictorial representation

Approach Framework

- The problem requires in depth analysis of previous campaign's customer data, and compare it with background data of the current customers to find optimal 1000 customers to be contacted.
- So, the model to predict the optimal customers to be contacted is based on **Random Forest** which is a supervised learning algorithm that is based on the ensemble learning method .

Data Processing and Treatment

→ **Excel**

- The data provided consisted of few blank cells. To deal with that, first we removed the rows which had missing values among the non-numeric columns.
- For missing values of numeric columns, we substituted the blanks with median of the corresponding column, so that the values of the other factors which contribute to the predicting model do not get ignored.
- The columns with binary values(i.e Yes/No) were converted to 1's and 0's (i.e. 1 → Yes & 0 → No)

→ **Python**

- For columns consisting of categorical non-binary data having nominal values (e.g- "marital status" variable with the values: "single", "married" and "divorced") which do not have natural ordering, One Hot Encoding technique is implemented, which replaces the data with dummy variables.
- After One Hot Encoding is done, and 'n' columns are obtained, we drop one of the columns, because 'n-1' columns provide sufficient information to determine that particular characteristic.

Model Selection

Reasons for model selection

- High accuracy
- The algorithm scales well when new features or samples are added to the dataset

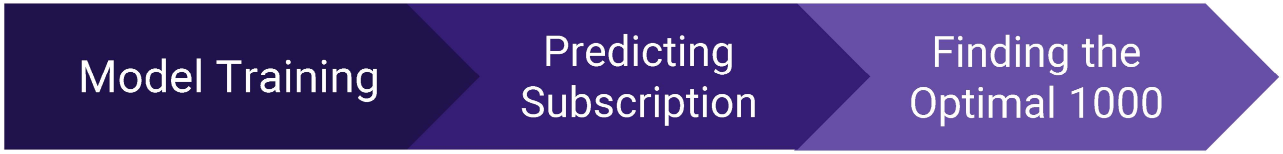
Random Forest is generally used if:

- It is not a time series problem
- The data has a non-linear trend and extrapolation is not crucial

Important parameters of model

R2 Score→0.8844552588169639
Mean Absolute Error→0.0572430262
Mean Square Error→0.01097271766
Root Mean Square Error→0.1047507

Methodology and Solution Design



- After processing the data, model is trained using Random forest algorithm, using historical data of previous campaign.
- It is applied because there is no linearity in the data set and the given problem is a binary classification problem.

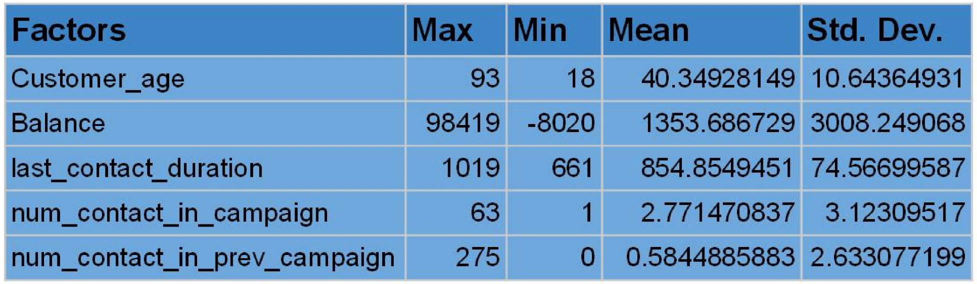
- The new customer background data is fed into the trained model.
- The predicted probabilistic subscription values for the targeted set of customers are obtained.

- The customer id and predicted values are exported to an excel file.
- The data is sorted in decreasing order of the predicted value.
- The top 1000 are the required customers.

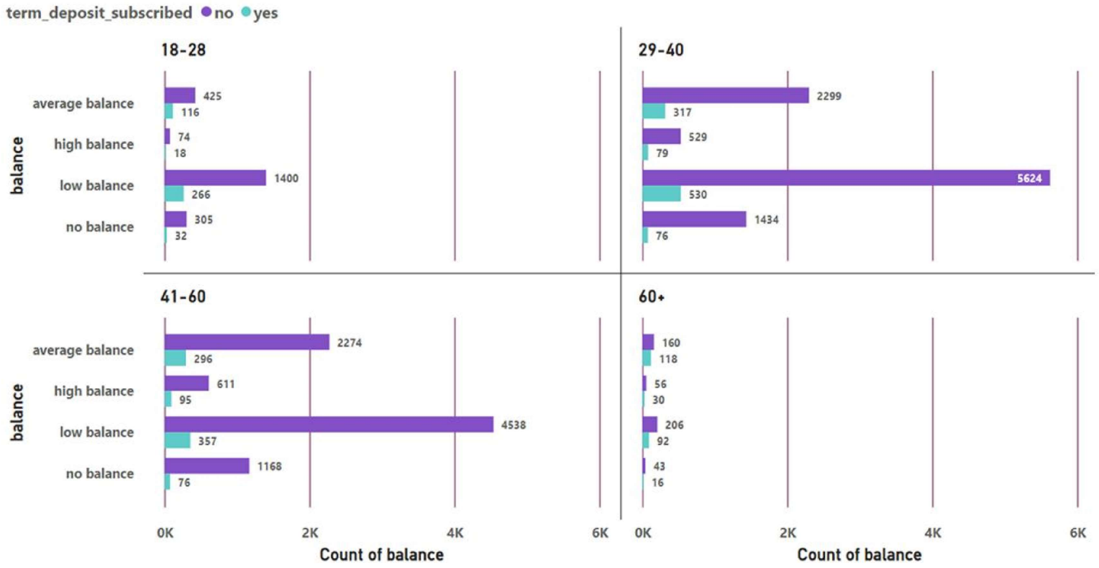
**The complete code for solution design is attached here in pdf format: PDF

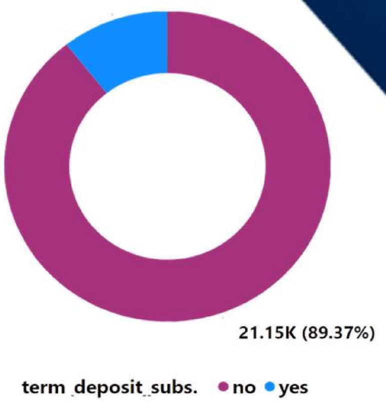
Code_Elite Sapiens.pdf

Data Visualisation



Factors	Max	Min	Mean	Std. Dev.
Customer_age	93	18	40.34928149	10.64364931
Balance	98419	-8020	1353.686729	3008.249068
last_contact_duration	1019	661	854.8549451	74.56699587
num_contact_in_campaign	63	1	2.771470837	3.12309517
num_contact_in_prev_campaign	275	0	0.5844885883	2.633077199





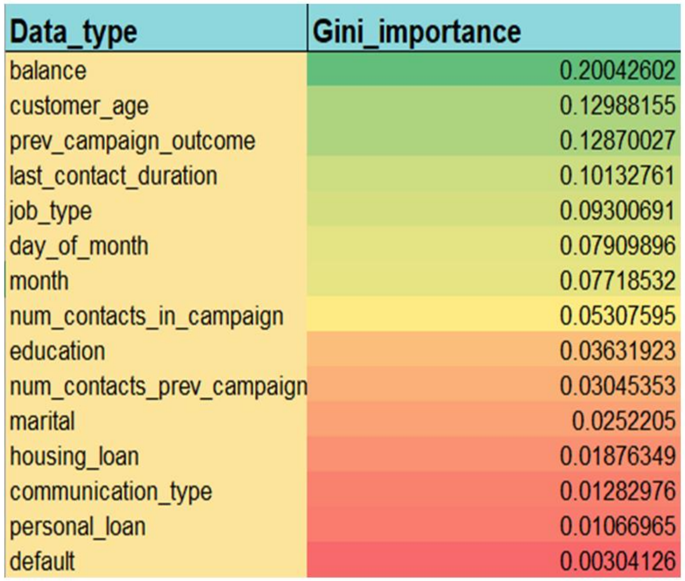
- no balance-(<0), low balance (1-1000), avg balance (1001-5000). high balance (>5000)
- Most number of the customers lie in the low balance category.
- Most number of customers lie in the age group of 29-40 while, least number of customers lie in 60+ age.
- Through the bar graph, conclusions can be drawn that 60+ age group in the average balance category has best conversion rate of taking term deposit in all balance categories and 29-40 age group in average balance category has the least.

**The complete data visualisation charts are attached here:

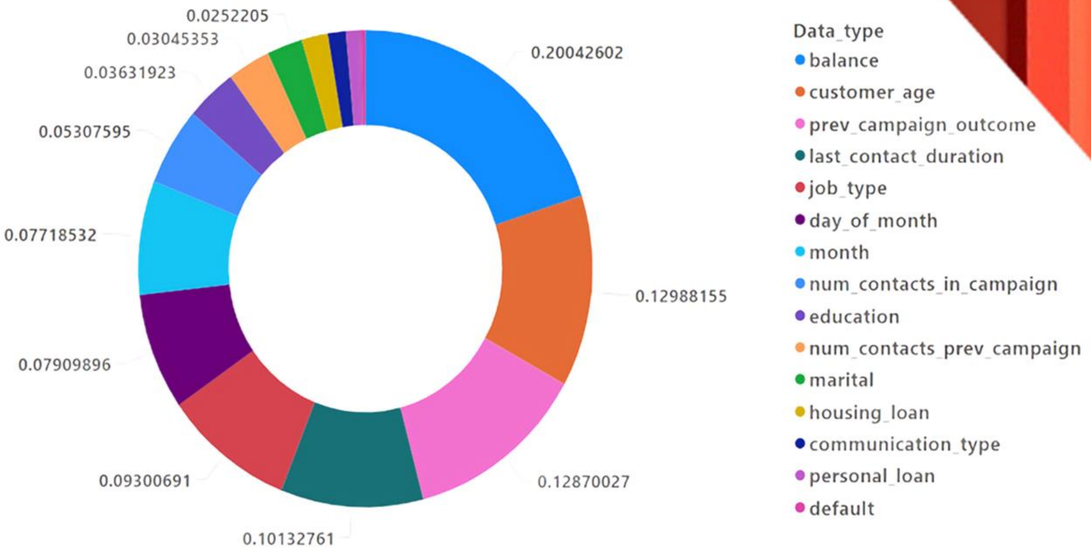
Data Visualisation_Elite Sapiens.pdf

Factors that contributed the most in Subscription

The importance of a feature is computed as the (normalized) total reduction of the criterion brought by that feature. It is also known as the Gini importance.



Data_type	Gini_importance
balance	0.20042602
customer_age	0.12988155
prev_campaign_outcome	0.12870027
last_contact_duration	0.10132761
job_type	0.09300691
day_of_month	0.07909896
month	0.07718532
num_contacts_in_campaign	0.05307595
education	0.03631923
num_contacts_prev_campaign	0.03045353
marital	0.0252205
housing_loan	0.01876349
communication_type	0.01282976
personal_loan	0.01066965
default	0.00304126


Data_type
balance
customer_age
prev_campaign_outcome
last_contact_duration
job_type
day_of_month
month
num_contacts_in_campaign
education
num_contacts_prev_campaign
marital
housing_loan
communication_type
personal_loan
default