Case Study

**Steps:**

- Category features are transformed to integers starting from 0 with sklearn.preprocessing
- LightGBM is chosen for this classification task, it stands for lightweight gradient boosting machines. It is more accurate and faster than other decision tree-based models.
- Since I chose integer encoding for categorical features, I wanted to try decision trees. I could use other encoding methods such as one-hot encoding frequency encoding etc., and try linear models but for simplicity, I inclined to decision tree models.
- 5 fold cross-validation is performed on the data. In each iteration, 4 fold is taken as the training set and 1 fold is as the validation set.

Model Performance:

This is an imbalanced dataset with '0: 37104, 1: 2896' so we need to check f1-score, presion, recall and ROC curve rather than accuracy. Also, a confusion matrix is presented for the overview of the result.

LightGBM: ROC AUC=0.950
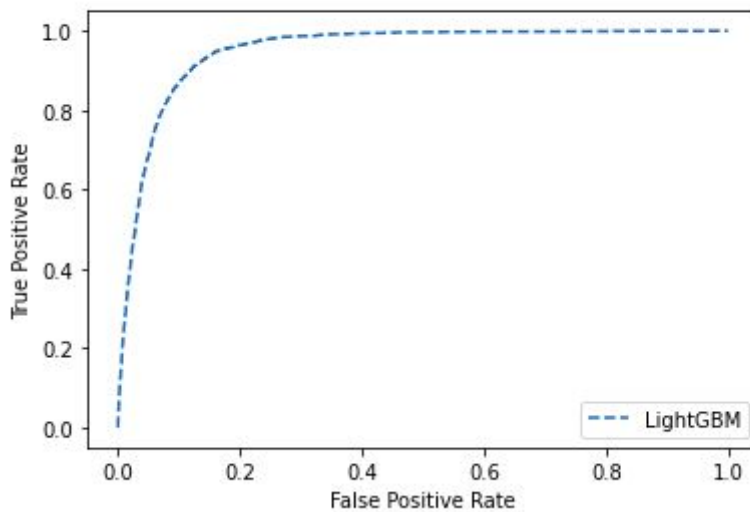
accuracy: 0.9382
recall: 0.41816298342541436
precision: 0.6061061061061062

|	Predicted 0	Predicted 1
Actual 0	TN:36317	FP:787
Actual 1	FN: 1685	TP: 1211

Feature Importance:
Feature importance is checked. It is seen that 'duration' (last contact duration) is most correlated feature with y(has the client subscribed to a term deposit) we should focus on 'duration' feature.'Balance', 'month' and 'day' come after 'duration'. When customers are in contact with the bank they are more likely to buy the product.
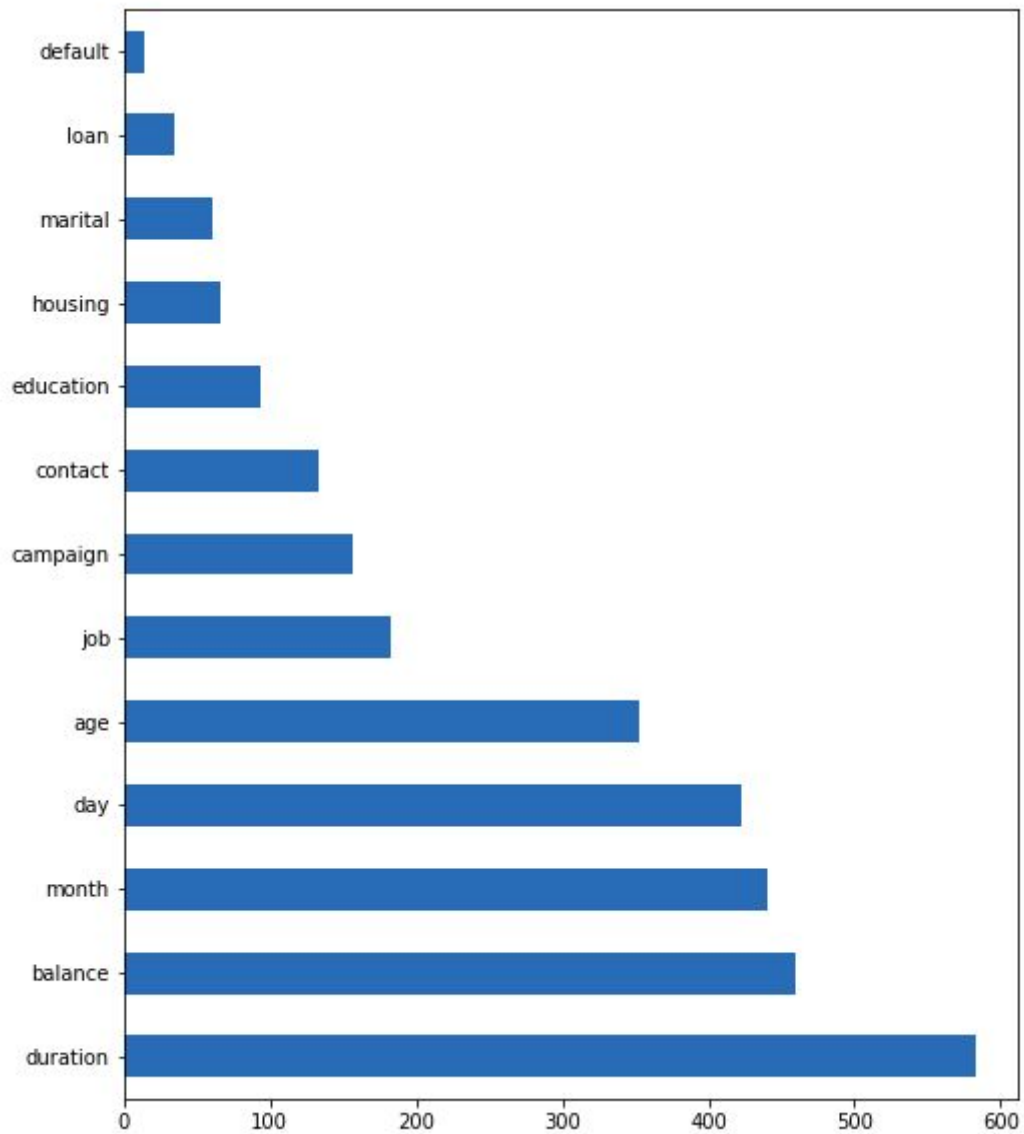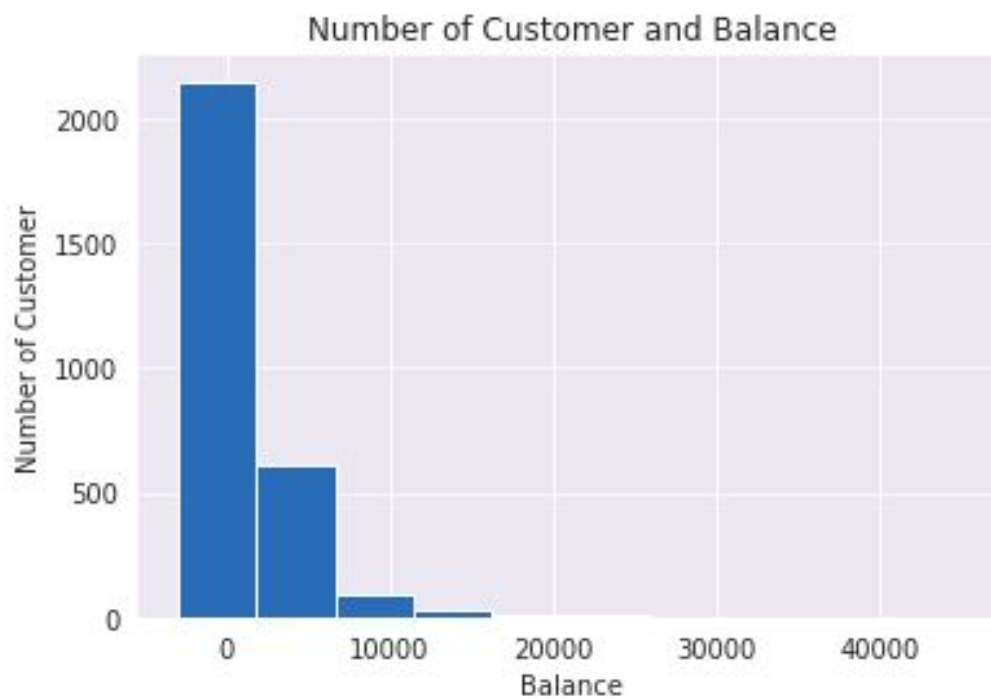
Customers that should be prioritized:

Customers are grouped based on y(has the client subscribed to a term deposit). Clients that have 'management' and 'blue-collar' are more likely to have a term deposit.
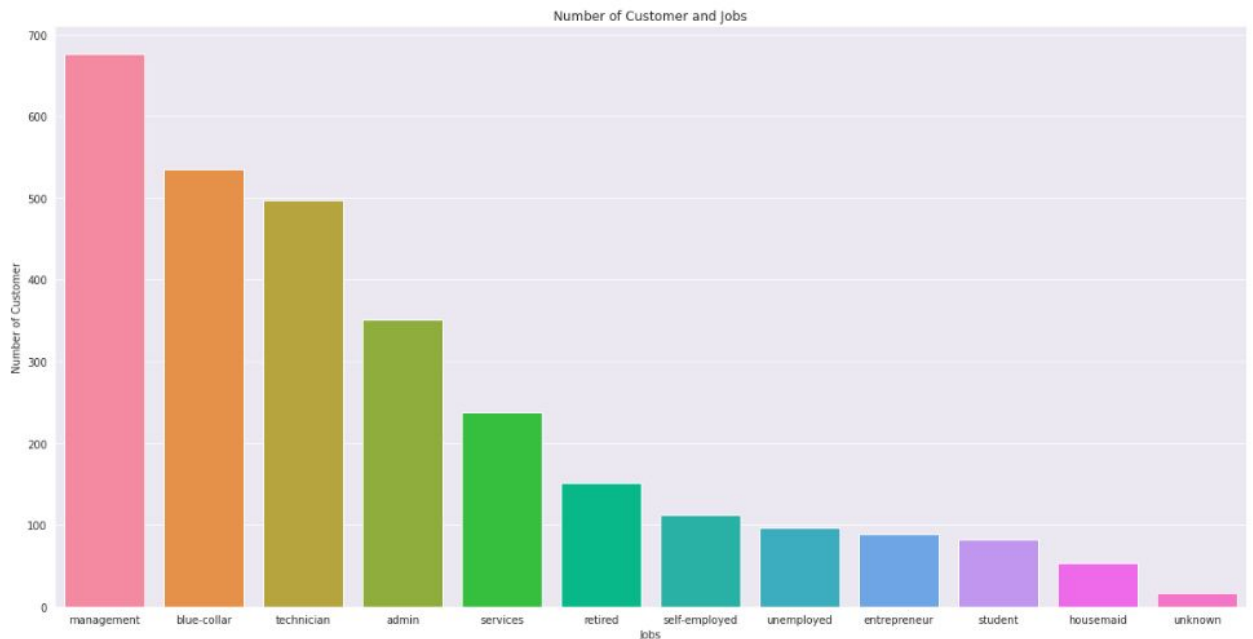
Customers that have low balance are more likely to have a term deposit.

Customers that are around 30 year old are more likely to have a term deposit.

Number of Customer and Ages
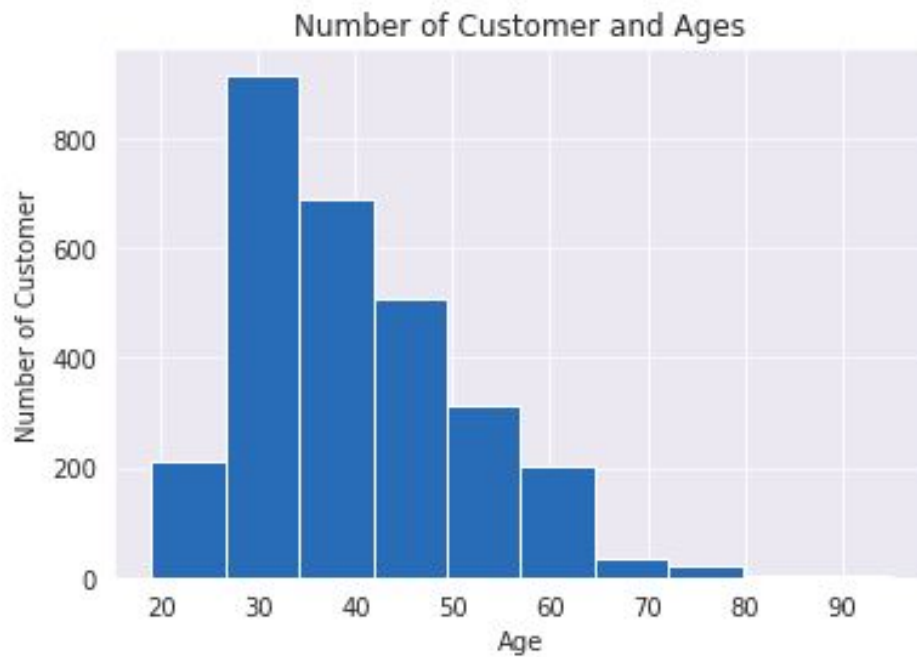
Since 'Month' feature has high importance in the model, it is also checked and seen that in May clients are more likely to subscribed to a term deposit

Number of Customer and Month