Bank Customer Churn Analysis

Using Machine Learning

Introduction

Presenter Profile	Rommel Labastida - Data analyst
Organization	YouSave Bank
Target audience	Marketing teams
Business case	How to keep current customers from churning

Problem Statement

“Predict bank customer churn to save potential active customers from churning

Objectives

- Identify which factors contribute the most to customer churn
- Build multiple prediction models to make sure the best model that can classify customer churn will not be missed

Exploratory Data Analysis

Exploratory Data Analysis: Data profiles

14 features with no missing values

	Column	Data Type	Unique Values	Null Values
0	RowNumber	int64	10000	0
1	CustomerId	int64	10000	0
2	Surname	object	2932	0
3	CreditScore	int64	460	0
4	Geography	object	3	0
5	Gender	object	2	0
6	Age	int64	70	0
7	Tenure	int64	11	0
8	Balance	float64	6382	0
9	NumOfProducts	int64	4	0
10	HasCrCard	int64	2	0
11	IsActiveMember	int64	2	0
12	EstimatedSalary	float64	9999	0
13	Exited	int64	2	0

10,000 records

```
df.shape
```

```
(10000, 14)
```

Irrelevant features

RowNumber	CustomerId	Surname
1	15634602	Hargrave
2	15647311	Hill
3	15619304	Onio
4	15701354	Boni
5	15737888	Mitchell

Sample rows of data

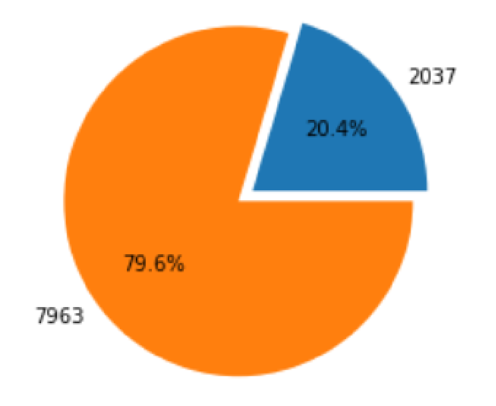
RowNumber	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
1	15634602	Hargrave	619	France	Female	42	2	0	1	1	1	101348.88	1
2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
3	15619304	Onio	502	France	Female	42	8	159660.8	3	1	0	113931.57	1
4	15701354	Boni	699	France	Female	39	1	0	2	0	0	93826.63	0
5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.1	0

'Exited' column == Churned
1 – Churned
0 – Not Churned

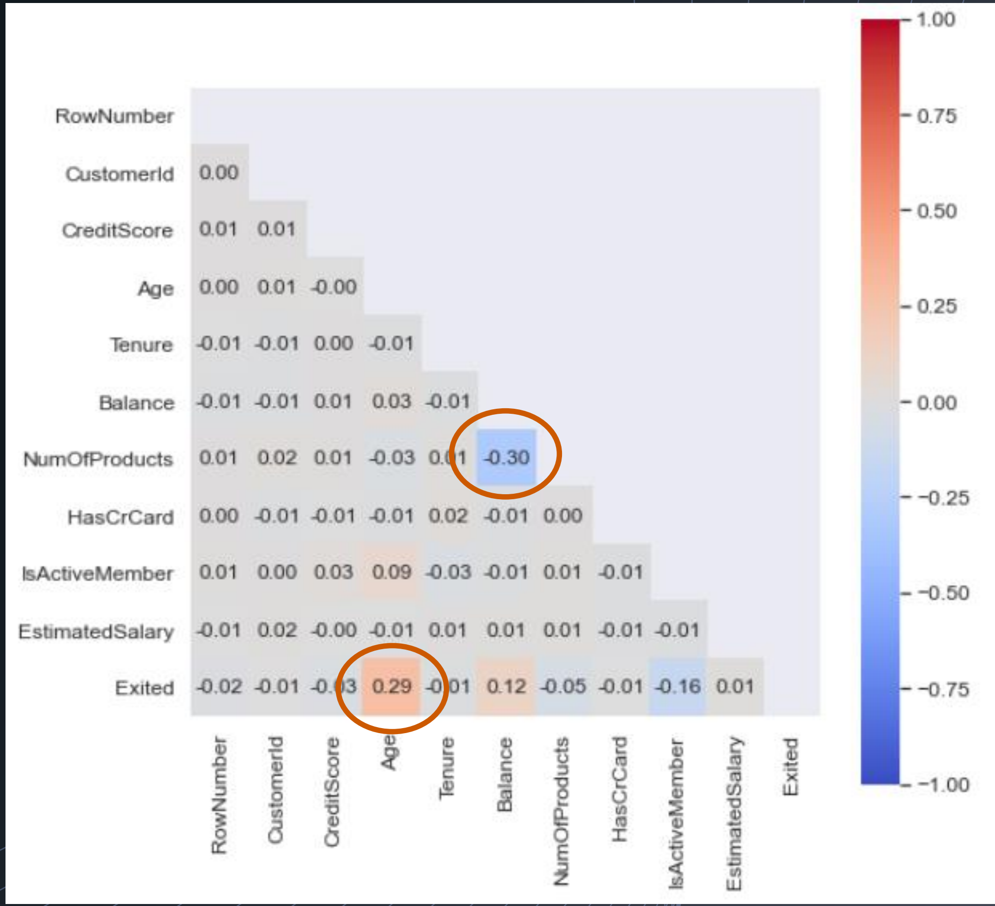
Data Source: Kaggle

Exploratory Data Analysis: Balance and correlation

Customer churn ratio



Correlation of features

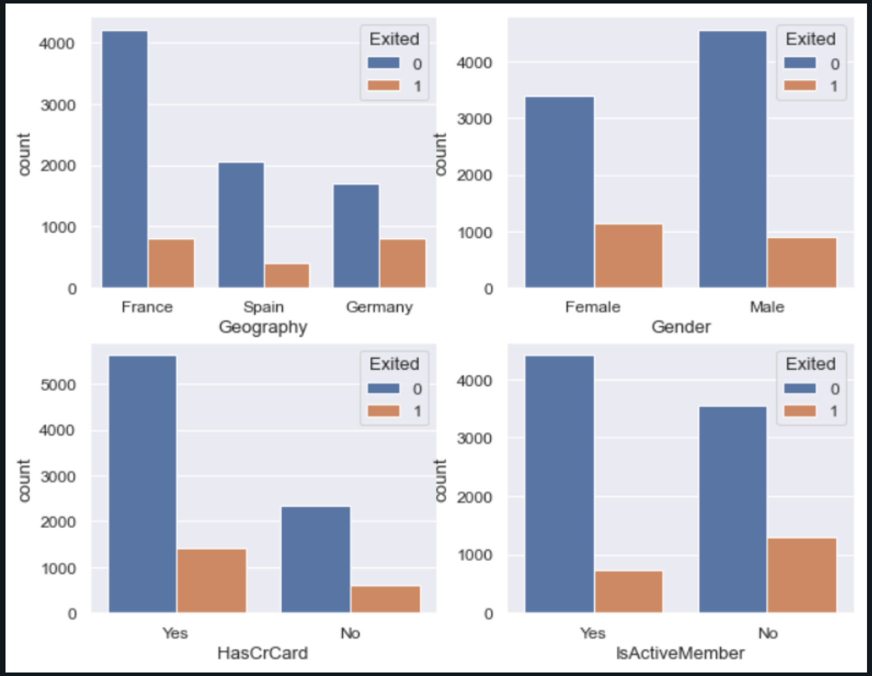


Exploratory Data Analysis: Relationship of features

Categorical Features

Geography

Gender



HasCrCard

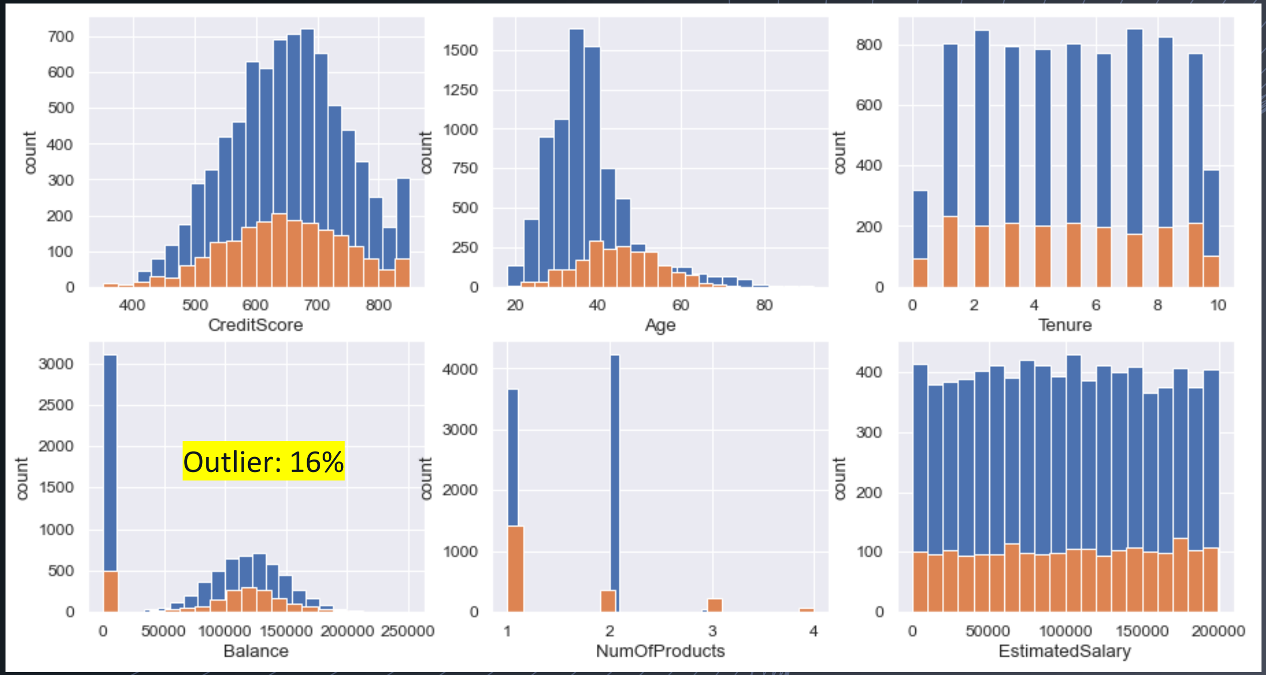
IsActiveMember

Continuous Features

CreditScore

Age

Tenure



Balance

NumOfProducts

EstimatedSalary

Churned: 500
Not Churned: 3117

1 – Churned
0 – Not Churned

Model Fit / Train / Test

Data Preparation and Model Fitting

Features encoding

	CreditScore	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited	Geography_France	Geography_Germany	Geography_Spain	Gender_Female	Gender_Male
0	619	42	2	0	1	1	1	101,348.88	1	1	-1	-1	1	-1
1	608	41	1	83,807.86	1	-1	1	112,542.58	0	-1	-1	1	1	-1
2	502	42	8	15,9660.8	3	1	-1	113,931.57	1	1	-1	-1	1	-1
3	699	39	1	0	2	-1	-1	93,826.63	0	1	-1	-1	1	-1
4	850	43	2	125,510.82	1	1	1	7,9084.1	0	-1	-1	1	1	-1

Using 30% test size

```
print(X_train.shape, y_train.shape)
print(X_test.shape, y_test.shape)
```

```
(7000, 13) (7000,)
(3000, 13) (3000,)
```

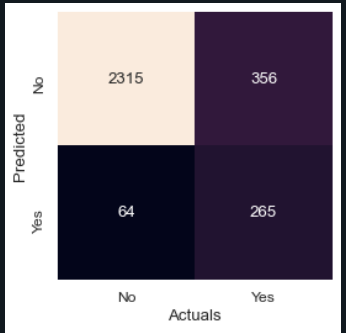
*SVM (Linear) 0.793

Models Scores: Default parameters

	Score	Confusion Matrix	Classification Scores	Features Importance
SVM	0.86	see below	see below	
Random Forest	0.867	see below	see below	see below
Logistic Regression	0.807	see below	see below	see below

SVM

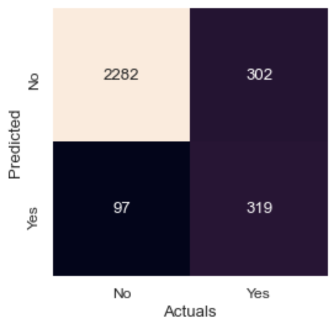
Score: 0.86



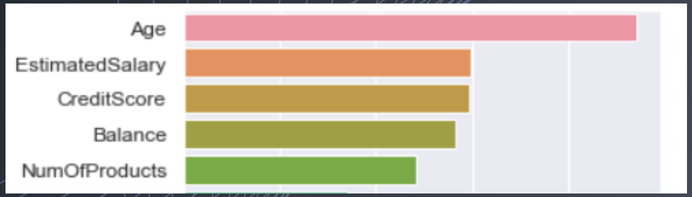
	precision	recall	f1-score	support
0	0.88	0.96	0.92	2379
1	0.77	0.51	0.62	621
accuracy			0.87	3000
macro avg	0.82	0.74	0.77	3000
weighted avg	0.86	0.87	0.86	3000

Random Forest

Score: 0.867

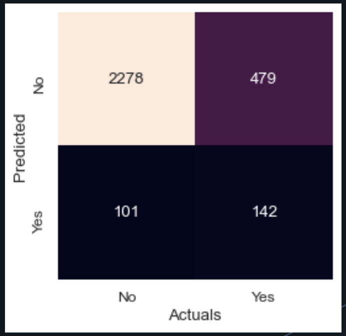


	precision	recall	f1-score	support
0	0.87	0.97	0.92	2379
1	0.81	0.43	0.56	621
accuracy			0.86	3000
macro avg	0.84	0.70	0.74	3000
weighted avg	0.85	0.86	0.84	3000



Logistic Regression

Score: 0.807



	precision	recall	f1-score	support
0	0.83	0.96	0.89	2379
1	0.58	0.23	0.33	621
accuracy			0.81	3000
macro avg	0.71	0.59	0.61	3000
weighted avg	0.78	0.81	0.77	3000

*SVM (Linear) Will be discarded as it has the lowest of 0.793

Cross Validation Scores: Baseline for 5-folds

	Mean Score	Standard Deviation
SVM	0.7054	0.0150
Random Forest	0.7401	0.0159
Logistic Regression	0.5954	0.0140

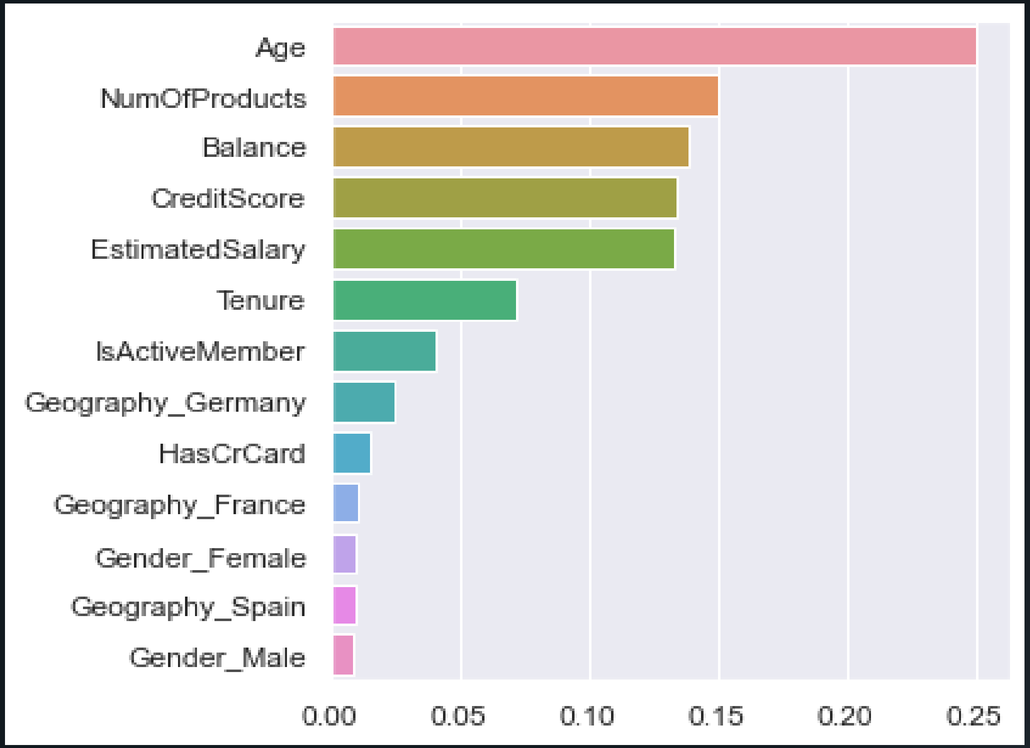
Random Forest: Best Estimator

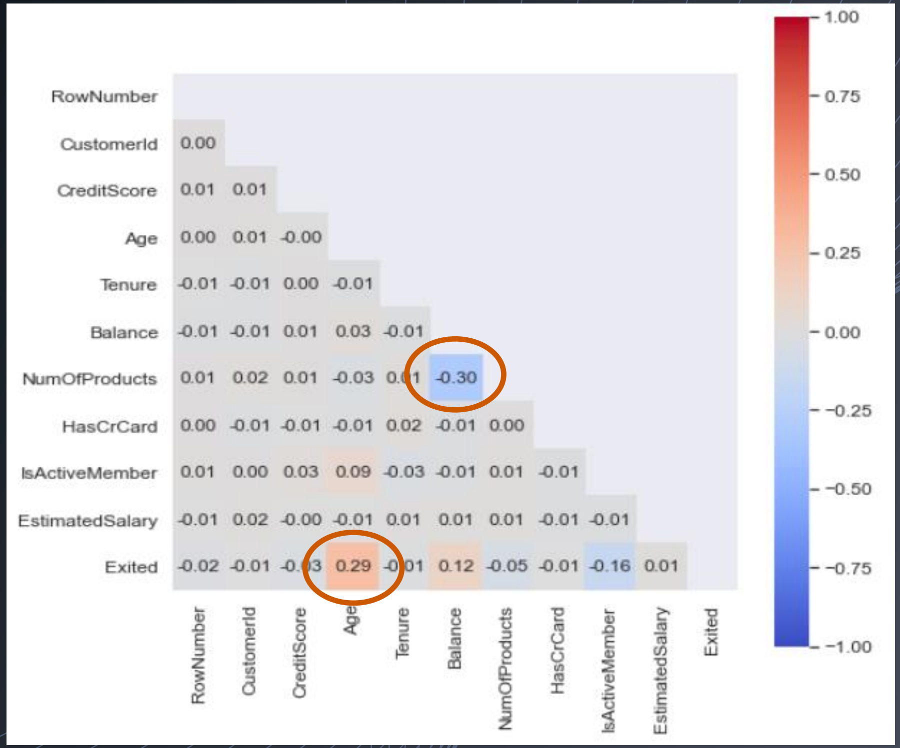
Best estimator	RandomForestClassifier criterion='entropy’, max_depth=25, min_samples_split=5, n_estimators=200
Best parameters	criterion: 'entropy’, max_depth: 25, min_samples_split: 5, n_estimators: 200
Best score	0.7443

Hyperparameter Tuning: Random Forest

	precision	recall	f1-score	support
Not Churned (0)	0.88	0.96	0.92	2379
Churned (1)	0.78	0.50	0.61	621
accuracy			0.87	3000
macro avg	0.83	0.73	0.77	3000
weighted avg	0.86	0.87	0.86	3000

Random Forest: Importance of features

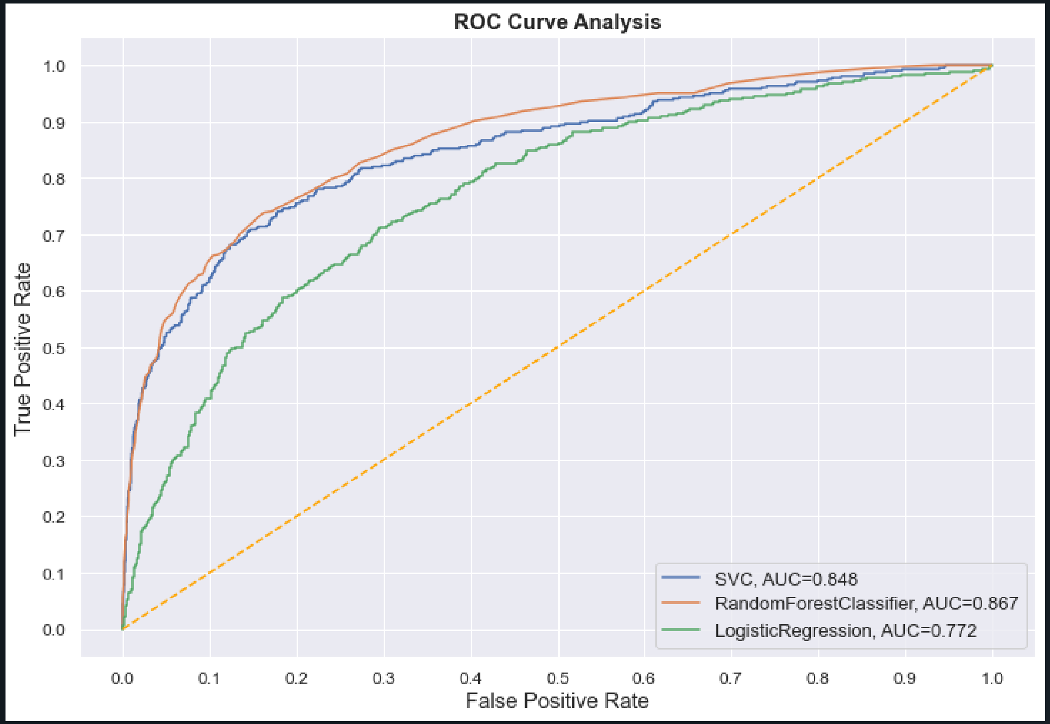




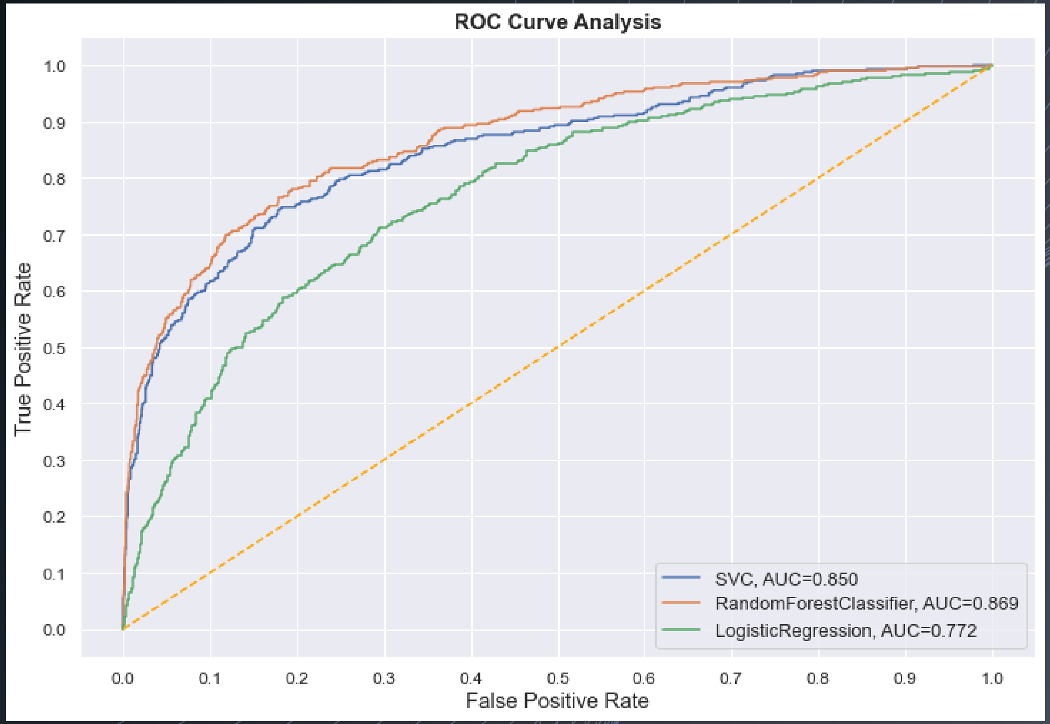
Random Forest model solidly confirms Age and NumOfProducts from the correlation heatmap earlier as contributing factors of customer churn

ROC Curve

Using default parameters



Using hyperparameters



Though there's not much gain in ROC curve using hyperparameter tuning, the ROC curves further confirms that Random Forest model gives a better balance between the precision (0.78) and recall (0.50) on 1's (exited)

Conclusions

“Though the Random Forest model scores accurately high at 87% and its precision in predicting who will churn in the test data is encouraging, the model was only able to catch 50% of those that eventually exited the bank.

Age is the clear determining factor to predict customer churn in the current data

Customers with exactly 2 bank products have very low tendency to churn compared to customers with 1 or more than 2

Recommendation

“The Random Forest model can be further improved by adding new features and more data as it currently suffers from imbalance target data at 20%

With the current data, the bank could meantime focus more on existing older customers to prevent them from churning (at least 50 % of them)

Give extra attention to bank customers with 1 bank product or more than 2 bank products as they have higher chance of churning

Thank you