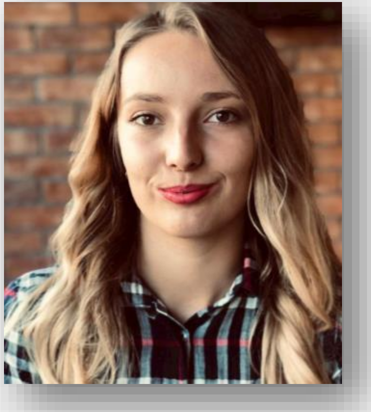
Santander Customer Transaction Prediction

Hello!

Shruti Bajpai



Alexandra Ionascu

Laura El Aoufir

1. Buisness Problem

Company overview

- Spanish multinational financial services company
- Madrid and Santander, Spain
- 16th largest banking institution in the world

Santander

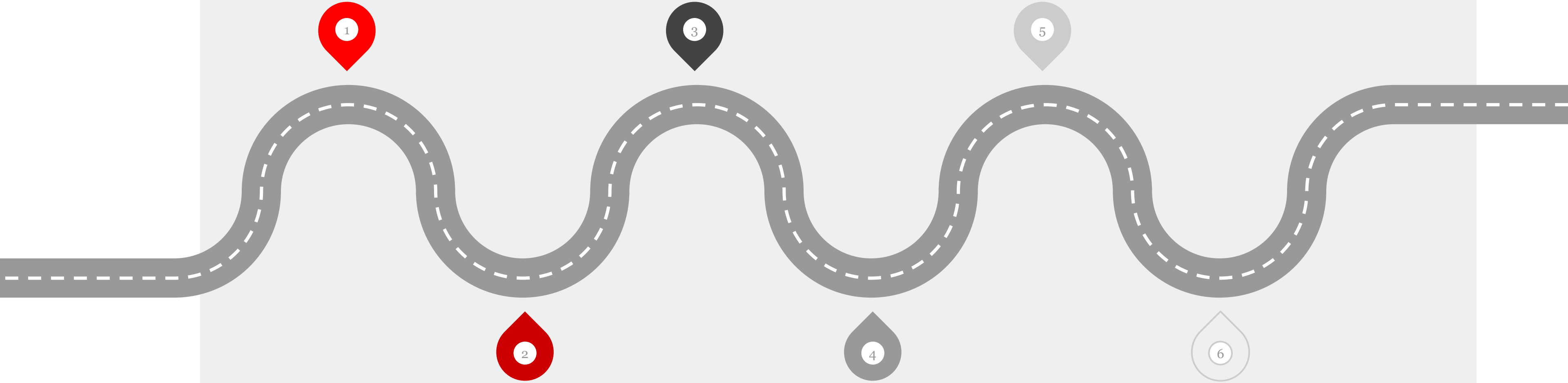
Data Science Objective

Classification, Supervised Learning

~

Identify which customers will make a specific transaction in the future, irrespective of the amount of money transacted.

Experiment overview



Data Exploration

Feature Selection
-
Pearson & Lasso

Oversampling
-
SMOTE

Modelling
Validation set approach
-
LR
DT
GB

Cross Validation
-
K-folds CV (no tuning)

Hyperparameter tuning
-
K-folds CV on DT

2. Data Exploration

Data Exploration

Shape

200k rows,

202 columns

Data types

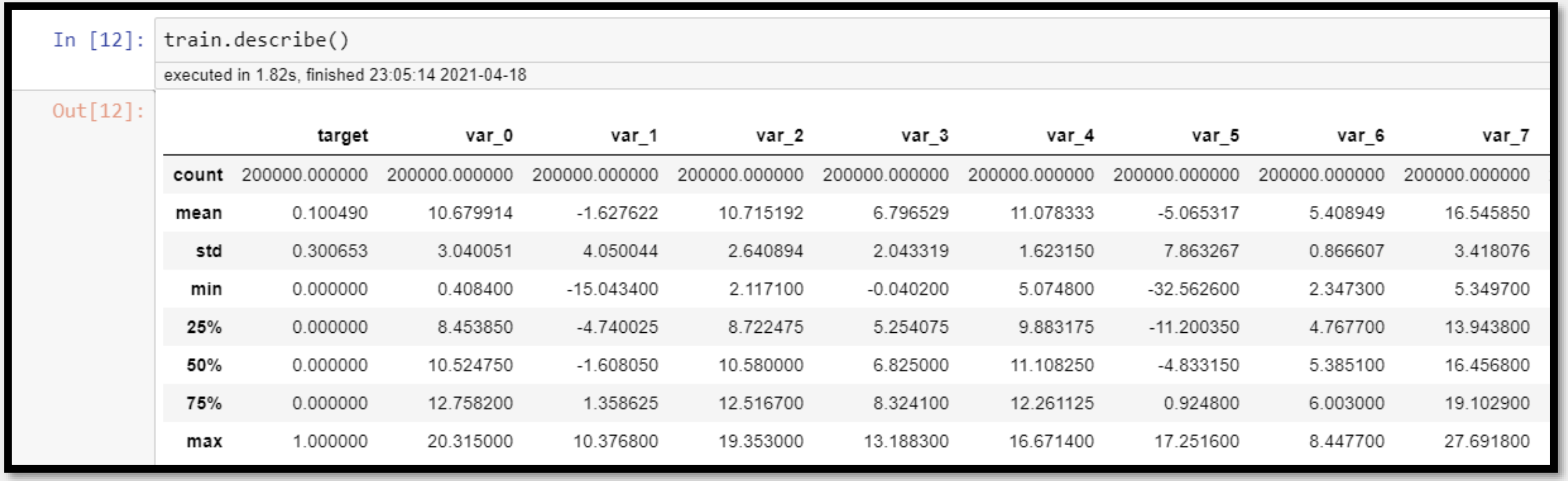
All numerical data

NAs

No missing values

Data

Standardized



```
train.describe()
```

executed in 1.82s, finished 23:05:14 2021-04-18

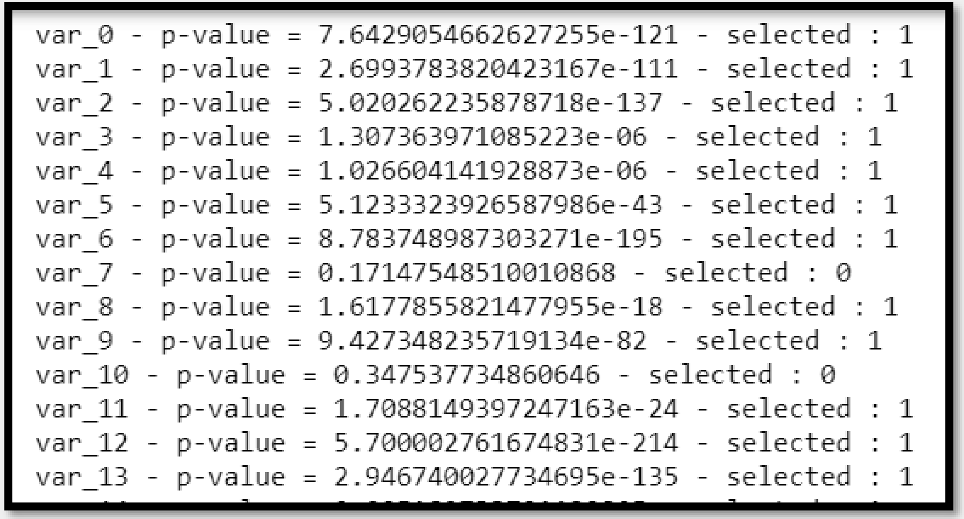
	target	var_0	var_1	var_2	var_3	var_4	var_5	var_6	var_7
count	200000.000000	200000.000000	200000.000000	200000.000000	200000.000000	200000.000000	200000.000000	200000.000000	200000.000000
mean	0.100490	10.679914	-1.627622	10.715192	6.796529	11.078333	-5.065317	5.408949	16.545850
std	0.300653	3.040051	4.050044	2.640894	2.043319	1.623150	7.863267	0.866607	3.418076
min	0.000000	0.408400	-15.043400	2.117100	-0.040200	5.074800	-32.562600	2.347300	5.349700
25%	0.000000	8.453850	-4.740025	8.722475	5.254075	9.883175	-11.200350	4.767700	13.943800
50%	0.000000	10.524750	-1.608050	10.580000	6.825000	11.108250	-4.833150	5.385100	16.456800
75%	0.000000	12.758200	1.358625	12.516700	8.324100	12.261125	0.924800	6.003000	19.102900
max	1.000000	20.315000	10.376800	19.353000	13.188300	16.671400	17.251600	8.447700	27.691800

3. Feature Selection

Pearson Correlation

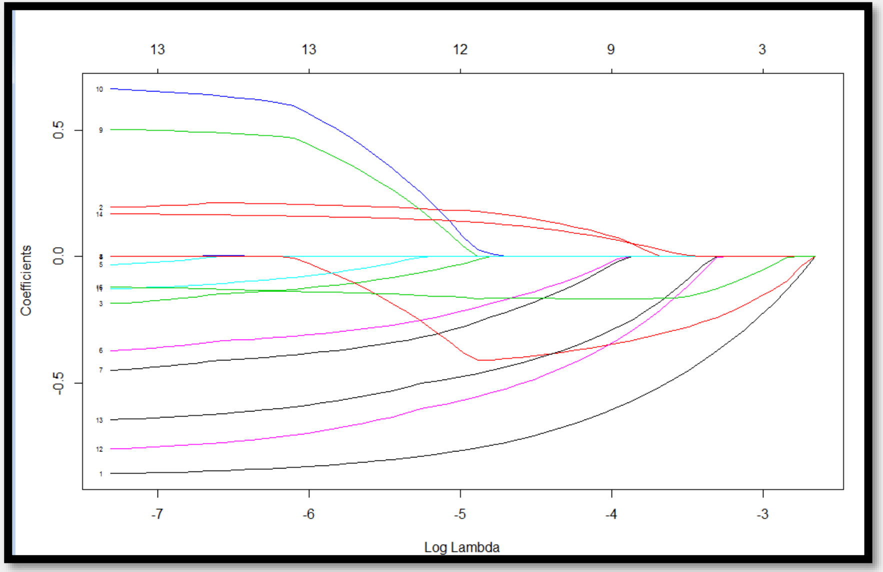
Selection Criteria: P-Value < 0.05

200 features → 181 features



```
var_0 - p-value = 7.642905466262725 5e-121 - selected : 1
```

Least Absolute Shrinkage & Selection Operator (LASSO)



200 features

↓

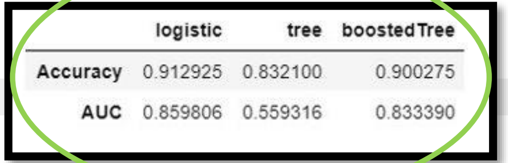
97 features

Methodology Comparison

Pearson Correlation

- Quick Analysis & Computation
- User can optimize selection of variables based on parameters; no such saturation exists

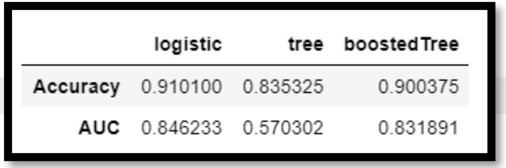
Method Selected



	logistic	tree	boostedTree
Accuracy	0.912925	0.832100	0.900275
AUC	0.859806	0.559316	0.833390

Lasso Regression

- Quick Analysis & Computation
- Selects at most ***n*** variables before it saturates
- Can eliminate variables that might increase the chances of higher prediction rates



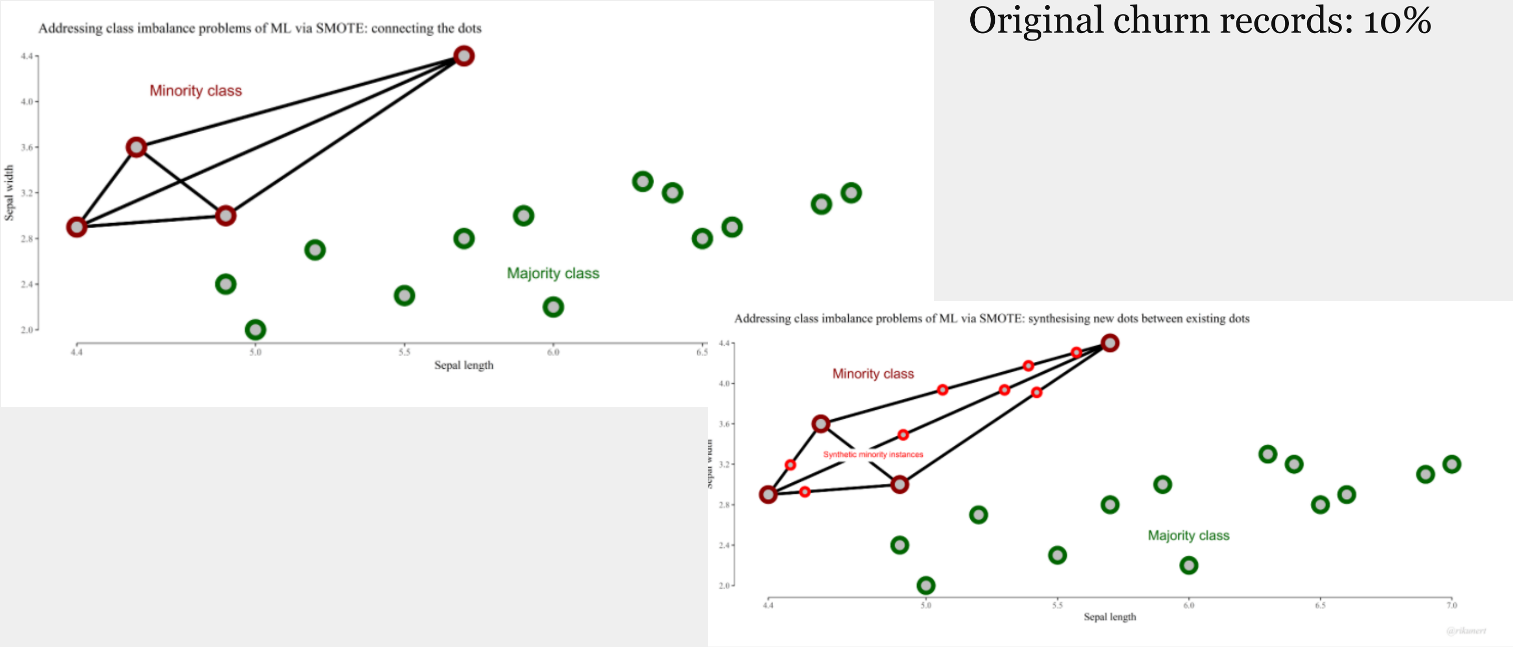
	logistic	tree	boostedTree
Accuracy	0.910100	0.835325	0.900375
AUC	0.846233	0.570302	0.831891

4. Oversampling

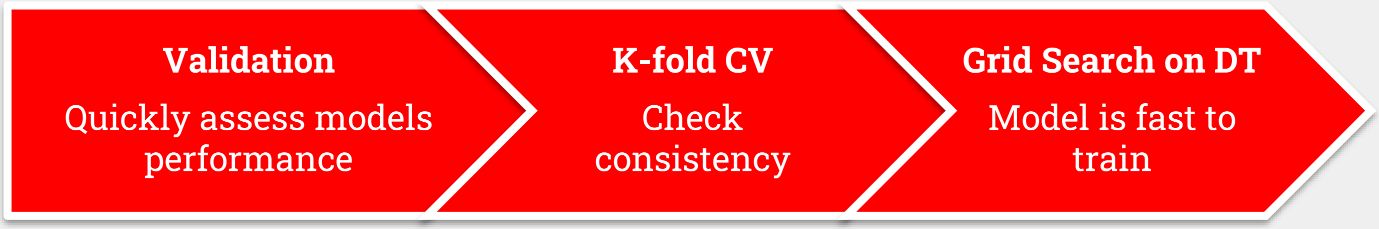
SMOTE

The algorithm creates new synthetic records between the real minority records

Original churn records: 10%



4. Modeling, Evaluation, Hyparameter Tuning



Validation

Quickly assess models performance

K-fold CV

Check consistency

Grid Search on DT

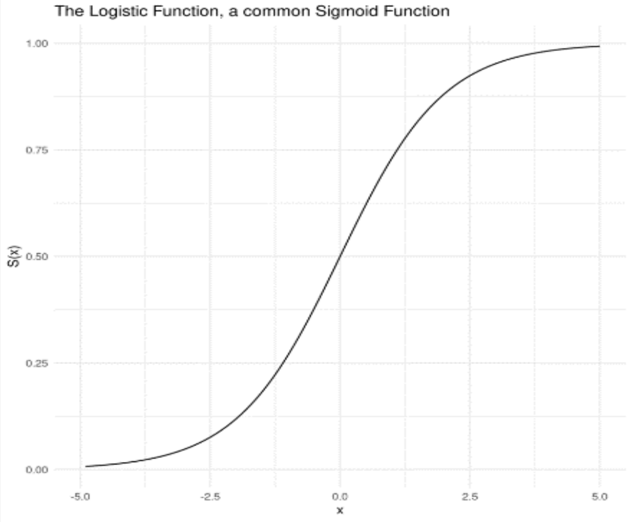
Model is fast to train

Modeling + Experimental setup

Logistic Regression

- Trained on data with oversampling
- **KAGGLE AUC Score: 0.77**

Metric Evaluation (Test_Split)	Score
AUC	0.8779
Accuracy	0.7988
KFOLD	0.798, 0.811, 0.795, 0.797, 0.796

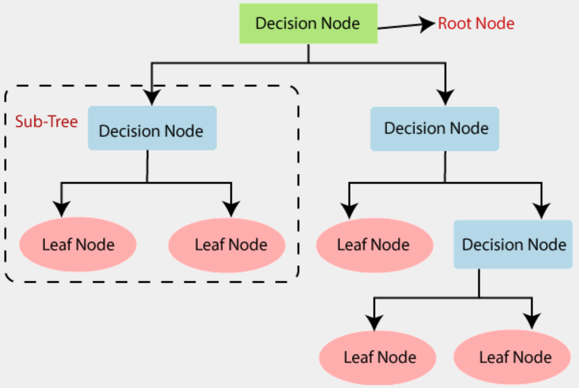


Modeling + Experimental setup

Decision Tree

- **Hyperparameter Training**: Grid Search on the max depth of the tree, using 5 folds
- Trained on data with oversampling
- **KAGGLE AUC Score: 0.56**

Metric Evaluation (Test_Split)	Score
AUC	0.56
Accuracy	0.833
KFOLD	0.548, 0.553, 0.553, 0.553, 0.550

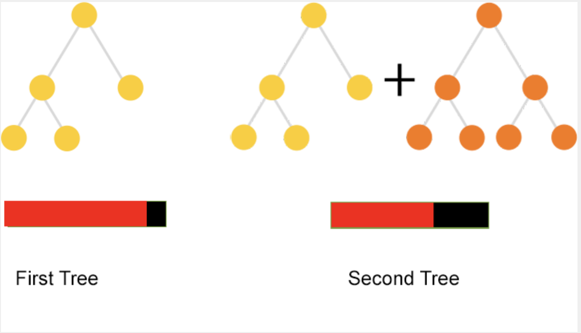


Modeling + Experimental setup

Gradient Boosting

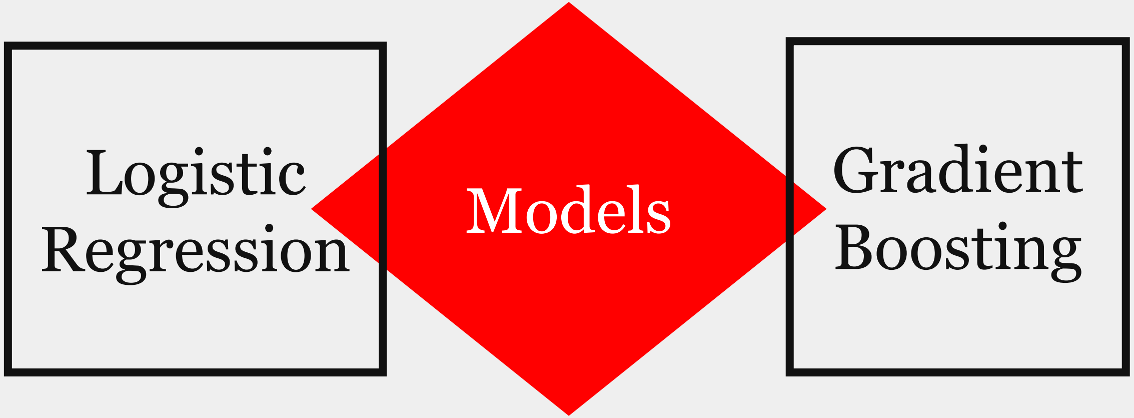
- **KAGGLE AUC Score: 0.8318**

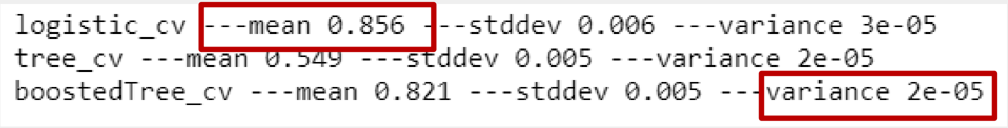
Metric Evaluation (Test_Split)	Score
AUC	0.83
Accuracy	0.90
KFOLD	0.818, 0.829, 0.821, 0.821, 0.815



★ Leaderboard position = 6756 / 8751

Best models





```
logistic_cv ---mean 0.856 ---stddev 0.006 ---variance 3e-05
tree_cv ---mean 0.549 ---stddev 0.005 ---variance 2e-05
boostedTree_cv ---mean 0.821 ---stddev 0.005 ---variance 2e-05
```

thanks!

Any questions?