# Santander Customer Satisfaction

# TABLE OF CONTENTS

#01 Business Problem and Impact

#02 Data Science Lifecycle

#03 Methodology

#04 Technological Stack

#05 Results

# BUSINESS PROBLEM

Santander

Santander Bank is interested in ***identifying dissatisfied customers*** early in their relationship.

They provide hundreds of ***anonymized features :***

*to predict if a customer is satisfied or dissatisfied with their banking experience.*

# BUSINESS IMPACT

People tend to trust ***advertising less and less***, and pay more attention to ***recommendations from people they trust*** — friends and people they know, or consumer ***opinions*** (reviews) they find online.

Poor customer experience/user engagement resulting in ***diminished customer loyalty***

***Inefficient and costly services.***

# COMMERCIAL

Unhappy customers will affect the ***churn rate*** but..

You had 5,000 customers on May 1
You had 4,920 customers on May 31st
Calculate churn rate: (5,000 – 4,920 )/5000 ~1% < 5% (ideal CR)

Pareto's handy little principle: ***20% of your customers generate 80% of your revenue.***

Identify your ***most valuable customers*** will help you decide a new strategy (paying attention, strength the relationships) and push up margins.

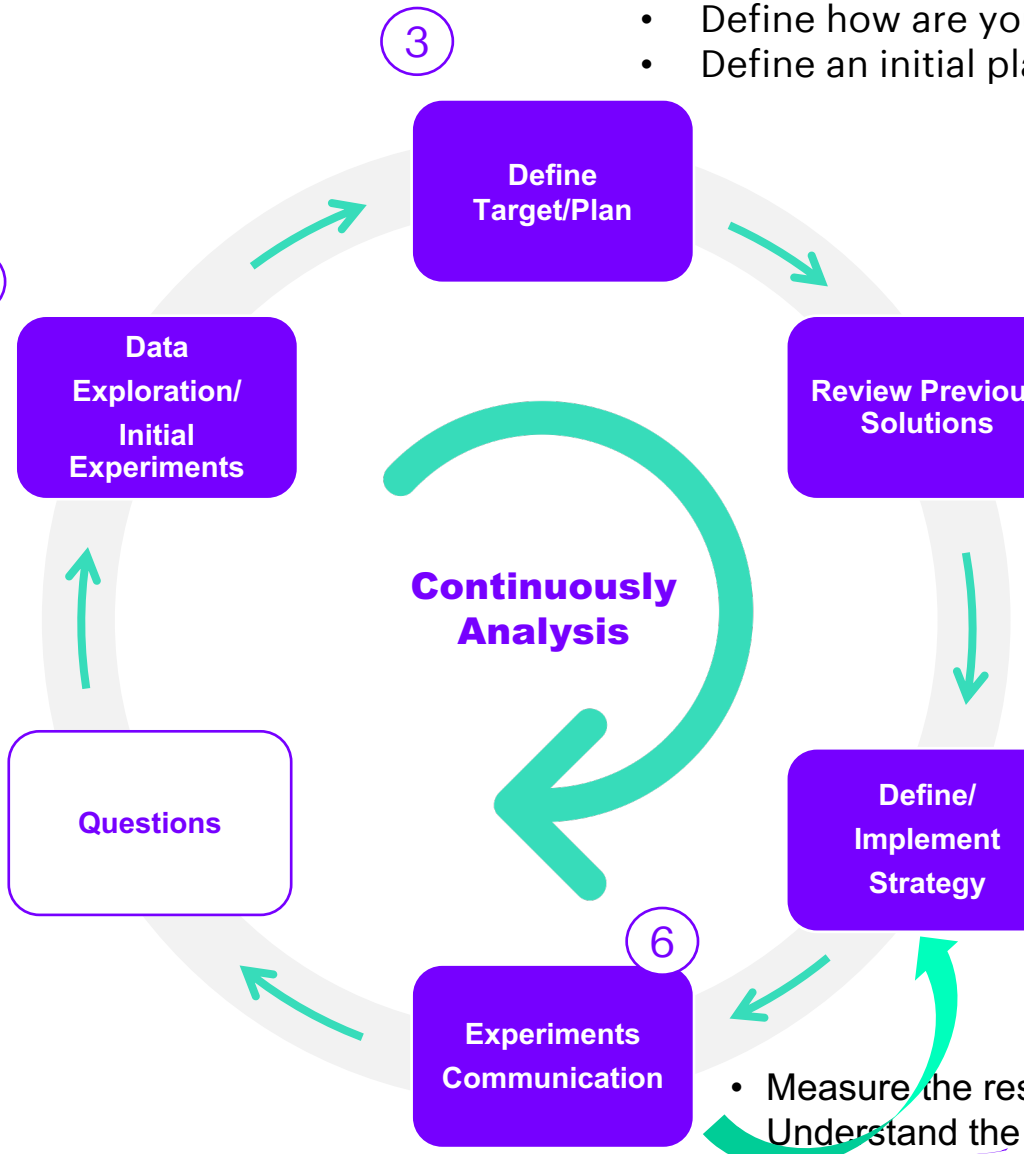
- 70% of customers left for the poor quality of the service (not product).
- 39% of consumers avoid vendors for over 2 years after having a negative experience.

# CUSTOMER CARE IN THE FUTURE

***Customer want a personalized, hyper-relevant experience***

We should figure out the needs ***(Predictive Care)*** of the customers and drive the  interaction with them to anticipate any issues and/or questions ***(Proactive Care)***

# DATA SCIENCE LIFE CYCLES




- Define a feasible target
- Define how are you going to measure it
- Define an initial plan

- Understand what you can reuse
- Where you should focus your attention
- Where you should innovate

- Define the Algorithms (Pipeline)
- Improve Implementation and Models
- Redefine the initial strategy

- Measure the results
- Understand the reasons behind the limits
- Compare respect to the target
- Communicate the Results/Limit
- Manage the Expectation

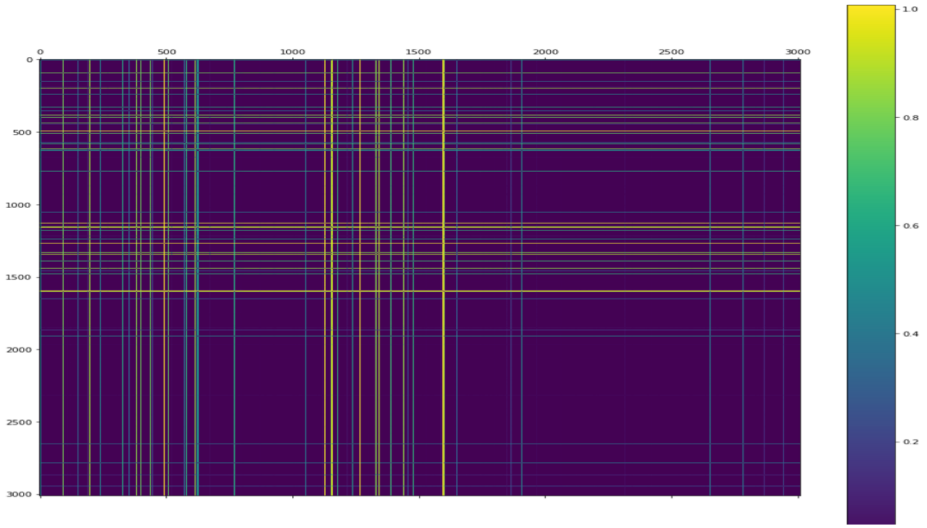
Initial Analysis of your data using
- visualization,
- sampling.

The main goal is to understand your data

- Formulate the questions, that you would like to answer with data ,
- Define the initial requirement (technological stack, legal, resources)

# DATA EXPLORATION

- Size of the dataset -> ***Scalability*** Problem
- Number of classes -> ***Unbalanced/Balanced*** Classification
- Presence of NaN values -> ***Clean*** Operations
- Initial Stats of the features -> Presence of ***Errors***
- Constant values among the features -> ***Features Elimination***
- Number of Features with binary values -> ***Sparse Features***




## Key Findings:

- **Clean**
  - ✓ Remove Constant Feature
  - ✓ Remove Duplicates.
  - ✓ Extreme Values in the Features
- **Feature Selection**
  - ✓ Separate binary and numeric feature
  - ✓ Understand their classification impact
- **Classification .**
  - ✓ Address the class-unbalanced problem
  - ✓ Define a classification strategy based
  - ✓ on feature processing and selection
  - ✓ Include clustering components.

# DEFINE TARGETS – INITIAL PLAN

The Leading Board solution is 0.82% in AUC with ~300 submissions, the target : 0.80% in AUC within week time ~35 hours

| Challenges | Solution | Time |
|---|---|---|
| Clean Data | Pandas Data Frame | ✓ 5 hours |
| Evaluate Presence of Extreme Values or Error | • Extreme Value Analysis<br>• Quantile Distribution<br>• Min-Max Analysis | ✓ 5 hours |
| Evaluate Impact of Binary Features | • Activation Analysis | ✓ 5 hours |
| Evaluate Impact of Real Features | • Measuring Improvement with RF<br>• Supervised Binning | ✓ 10 hours |
| Define Classification Pipeline | • Over-sampling/Under-sampling.<br>• Dimensionality Reduction<br>• Fine Turning Random Forest<br>• XGBoost | ✓ 10 hours |

# METHODOLOGY- STRATEGY (NUMERIC FEATURES)

1 - Divide ***Numeric/Binary Feature***

2- Define ***classification baseline*** with Random Forest → F1 accuracy: 0.541 (+/- 0.006)

3 - Check Presence of ***Error/Extreme Values***

4 - Define/Apply Strategy for ***removing Error/Extreme Values***

5 - Measuring ***Improvement*** respect to the base-line → F1 accuracy: 0.545 (+/- 0.005)

6 - Define/Apply a Strategy for ***Feature Discretization***

7 - Measuring ***Improvement*** respect to the base-line → F1 accuracy: 0.565 (+/- 0.006)

8 - Reduce Feature based on ***Feature Selection***

9 - Measuring ***Improvement*** respect to the base-line

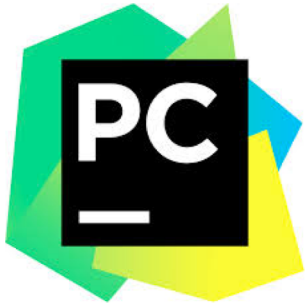
| | Predicted happy | Predicted unhappy |
|---|---|---|
| Real happy | 0.8670 | 0.1330 |
| Real unhappy | 0.5060 | 0.4940 |

# TECHNOLOGICAL STACK

***Data Analysis***

***Analytics Libraries***

***Uploading Prediction***


kaggle

NumPy

Pandas

scikit learn

matplotlib

open source

dmlc XGBoost

python virtualenv

python

Cython

***Efficiency***

***Data Analysis Code is not TDD-Code ☹!***

# RESULT (FINAL SUBMISSION-15)

1 – Remove ***Binary Feature***

2 - Remove ***Duplicate***

3 – Replace Feature with ***Extreme Values*** with their average value

4 –Normalize  Feature Values Using ***Quantile Transformation***

5 –Create Clusters Models based on ***K-Means***  on subsample

6 - Predict ***Cluster Labels*** for all sample using the K-Means model

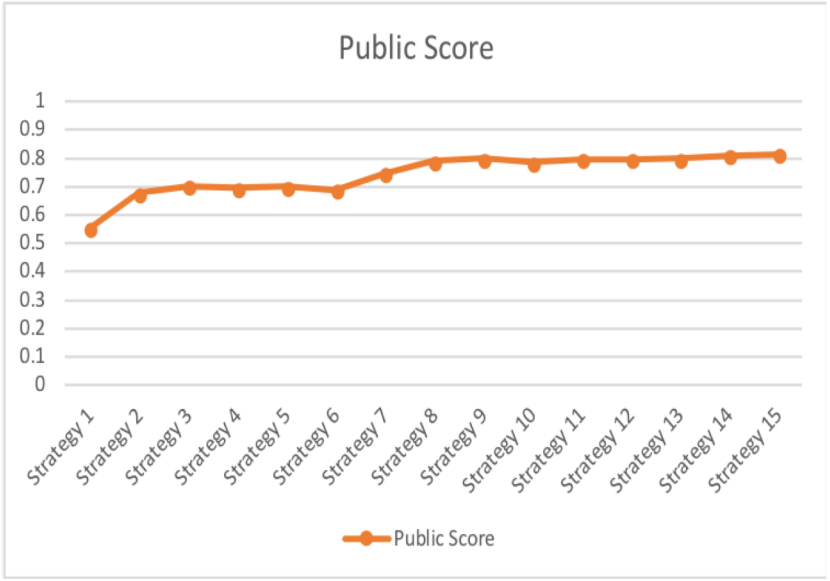
7 – Under-sampling the data using ***Tom-Link***

8 – Fine ***tuning an XG-Boost classifier*** for 400 rounds with a dev-set of 40%

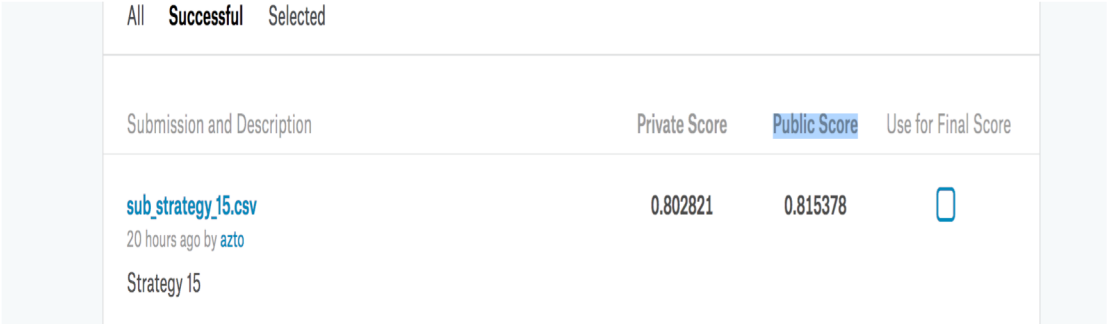
9 – Select the ***best XG-Boost model*** as train model

10 – Run the  Previous ***Transformation on Test Data***

11 – ***Prediction*** on test data




AVG Time ~  10 min



# WHAT SHOULD WE DO NEXT?

- **Feature Selections -> Remove Correlation and Improve Efficiency**
- **Ensemble/Stack Modelling-> Improve generalization (~ 2%-5%) – reduce efficiency , more parameters.**
- **Outlier prediction as new features, Improve the False Negative ( unhappy customers detected as happy)**
- **Explanation of the Prediction (LIME) -> Improve the Decision Making**

# Thank You