Santander
Product Recommendation

TEAM KWT Wen Li Yisong Tao Lydia Kan

01 Introduction

02 Data Cleaning and EDA

03 Feature Engineering

04 Training Model

05 Result and Finding

06 Future Steps

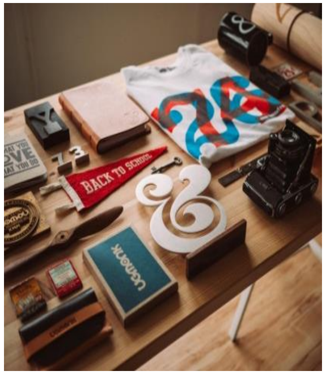
Introduction

Project Description

Santander Bank offers their customers personalized product recommendations time to time, in order to meet the individuals needs and satisfaction.
This challenge seeks to improve the recommendation system by predicting which products their existing customers will use in the next month based on their past behavior.

Goal

Achieve top 5% ranking and MAP@7 score on Kaggle leader board



Introduction

01

Data Size

Training Set:
13,647,409

Test Set: 929,615

02

Input Features

Categorical: 21
Continuous: 3

Customer Info. :
1: 24

03

Output Features

Product Purchased Info:
25:48

2015.1 – 2016.5

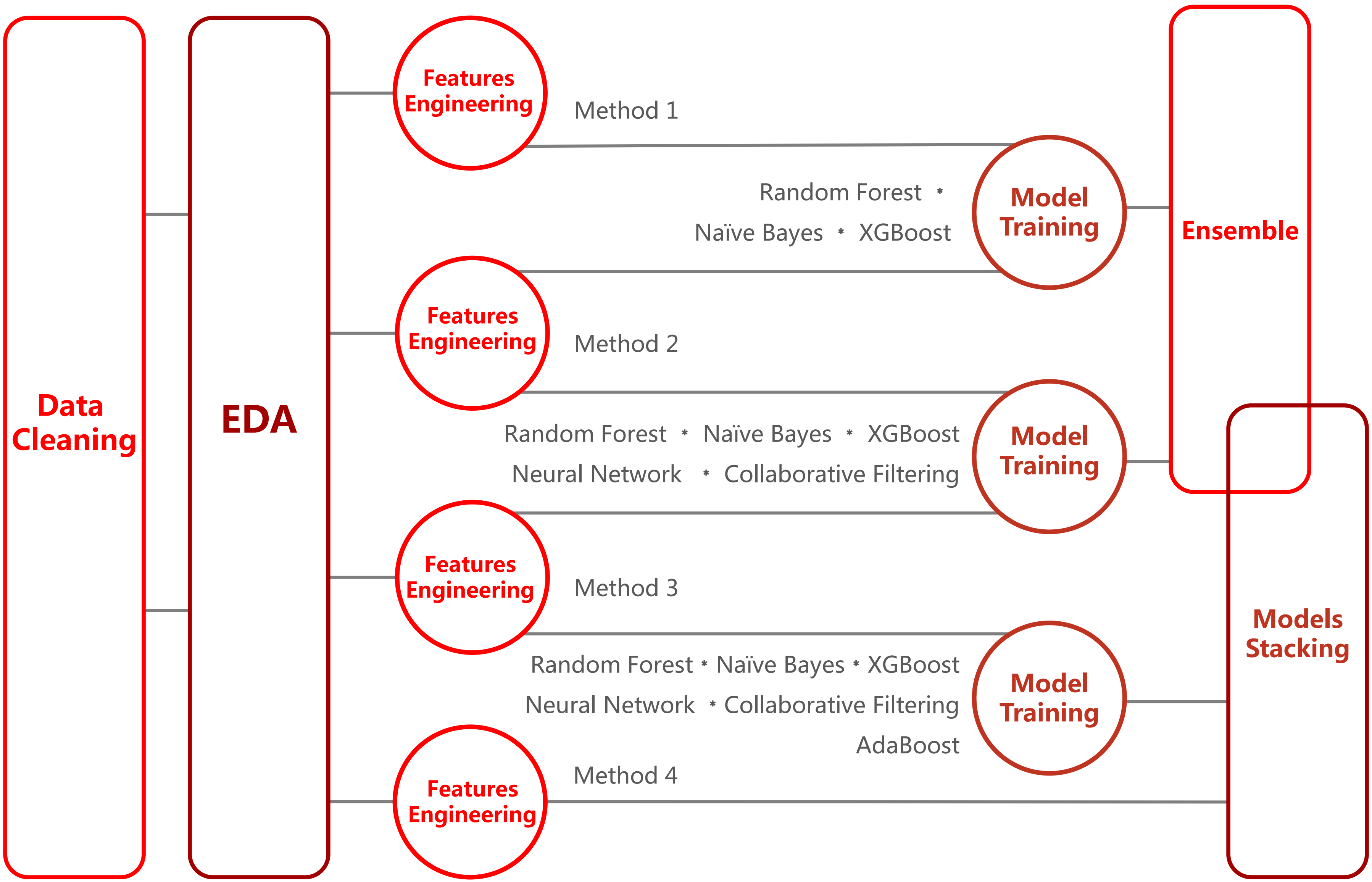
04

Evaluation

MAP@7

Multi-Classifier
Recommended
Products : 7

Workflow



Data Cleaning
EDA
Features Engineering
Method 1
Random Forest • Naïve Bayes • XGBoost
Model Training
Ensemble
Features Engineering
Method 2
Random Forest • Naïve Bayes • XGBoost
Neural Network • Collaborative Filtering
Model Training
Features Engineering
Method 3
Random Forest • Naïve Bayes • XGBoost
Neural Network • Collaborative Filtering
AdaBoost
Model Training
Models Stacking
Features Engineering
Method 4

Data Cleaning

Imputation

Contain Missing Values:

24 Features

Time Series – Customer Info.

Dropping Features

Drop 5 Features:

- **Having over 95% missing value**
- **Repetitive of other features**

Imputation

Unknown

- Sex
- Employee Index
- Country Residency
- Segmentation
- Residence Index
- Foreigner Index
- Channel to Join
- Primary
- Province Name

Common Type

- Customer Type
- Activity Index
- Income

Others

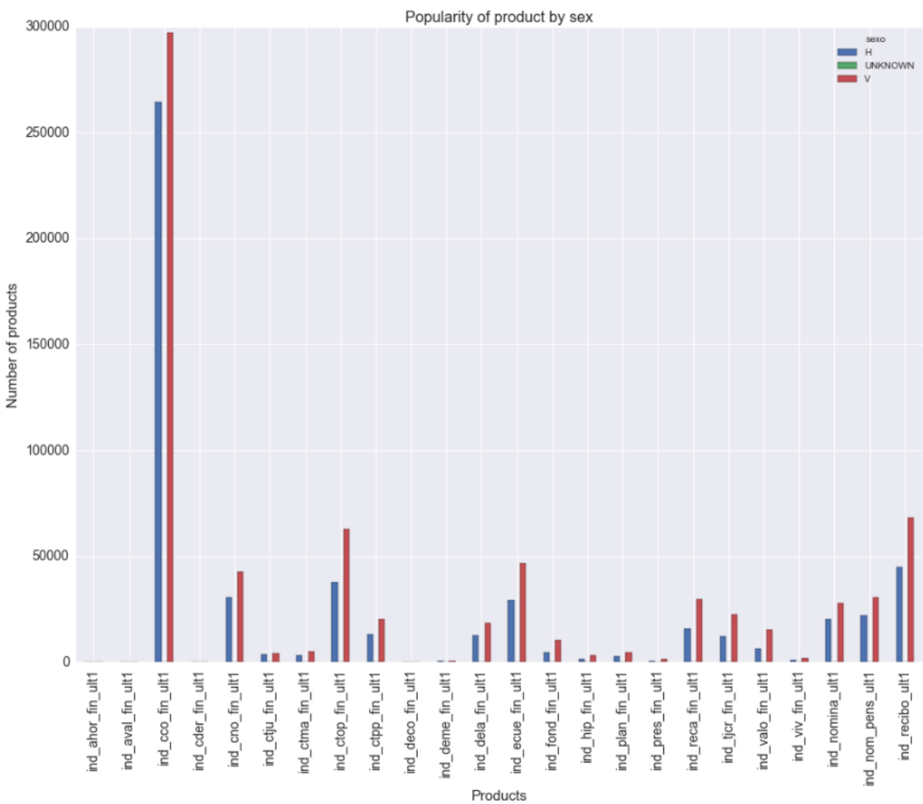
- New Customer – New
- Seniority – Min
- Age – Scale, Mean
- Relationship Type – 'A'
- Deceased Index – 'N'

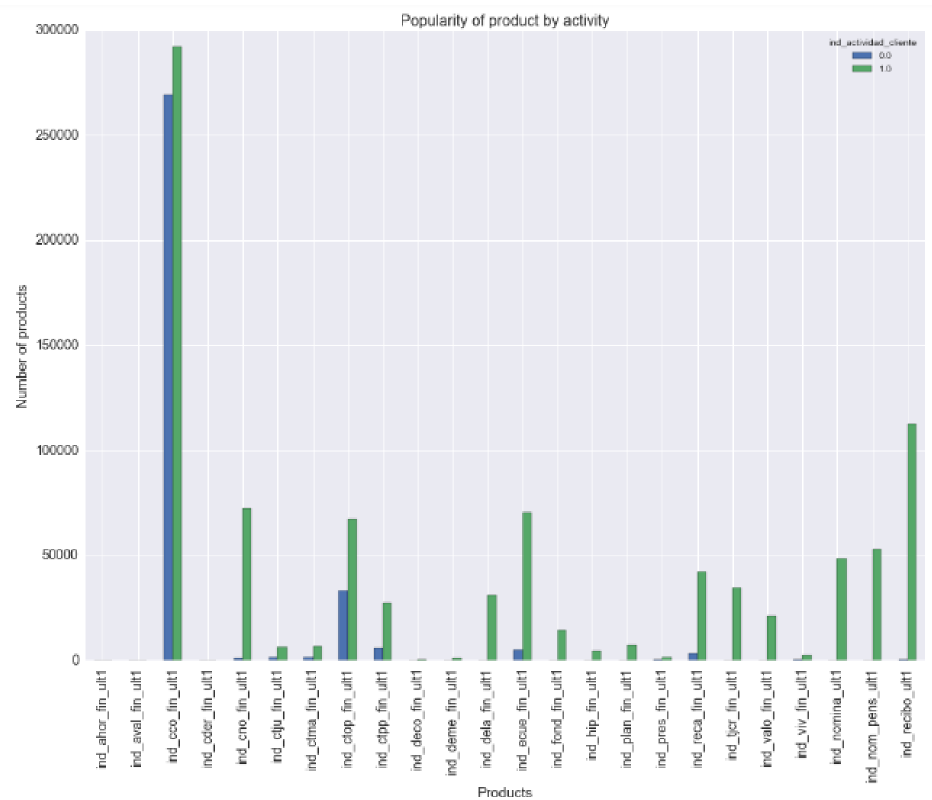
Products

- Payroll - 0
- Pensions - 0

EDA

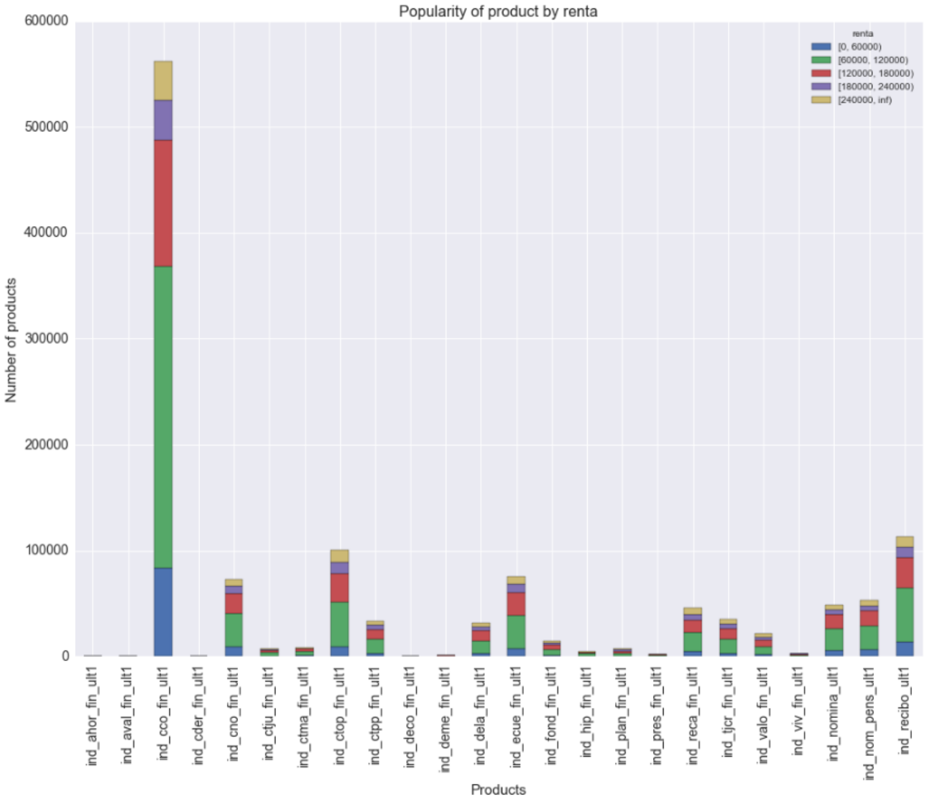
Product Sales Related to Customer's Info - 2016.5

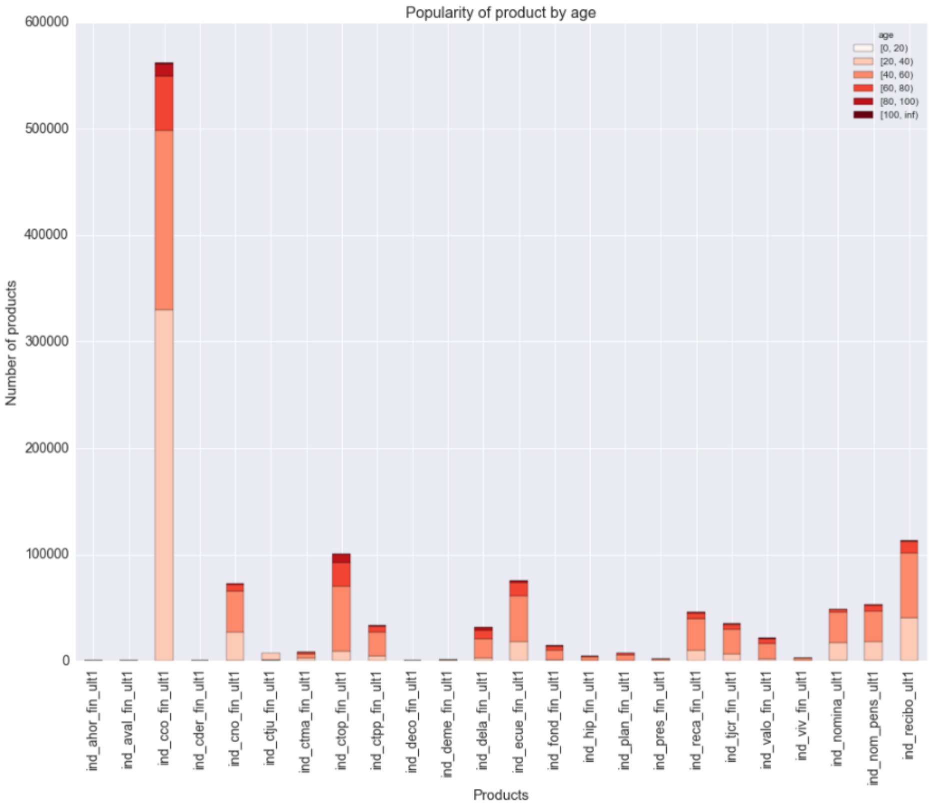




EDA

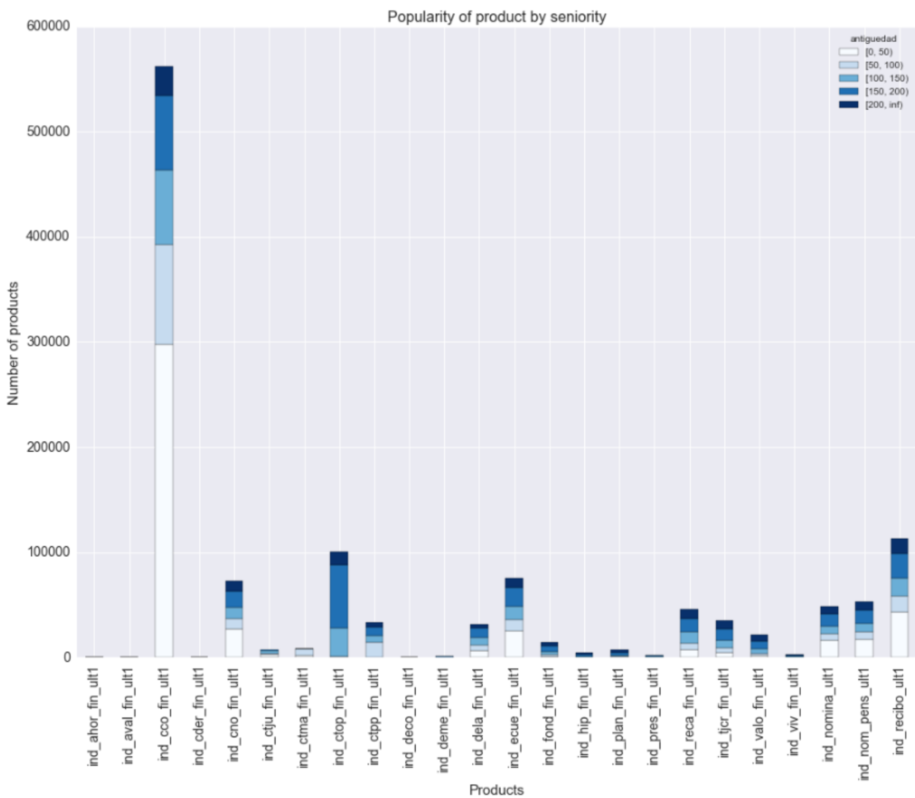
Product Sales Related to Customer’s Info - 2016.5

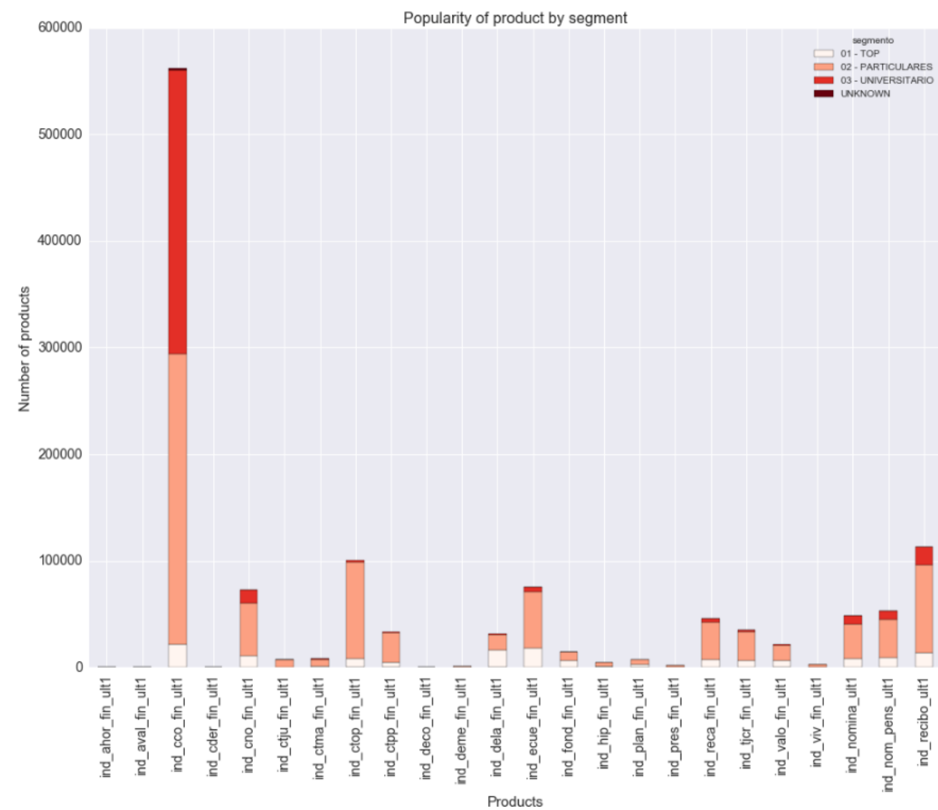




EDA

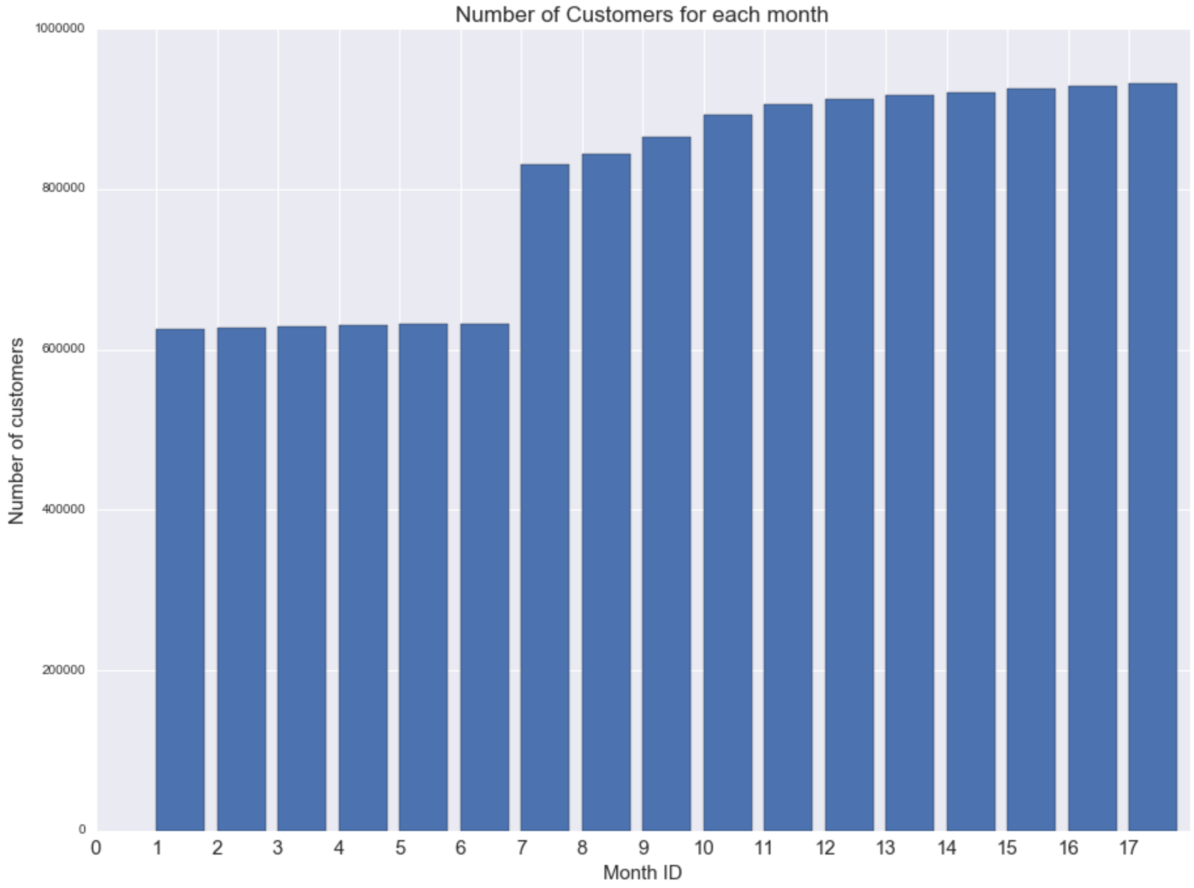
Product Sales Related to Customer’s Info - 2016.5





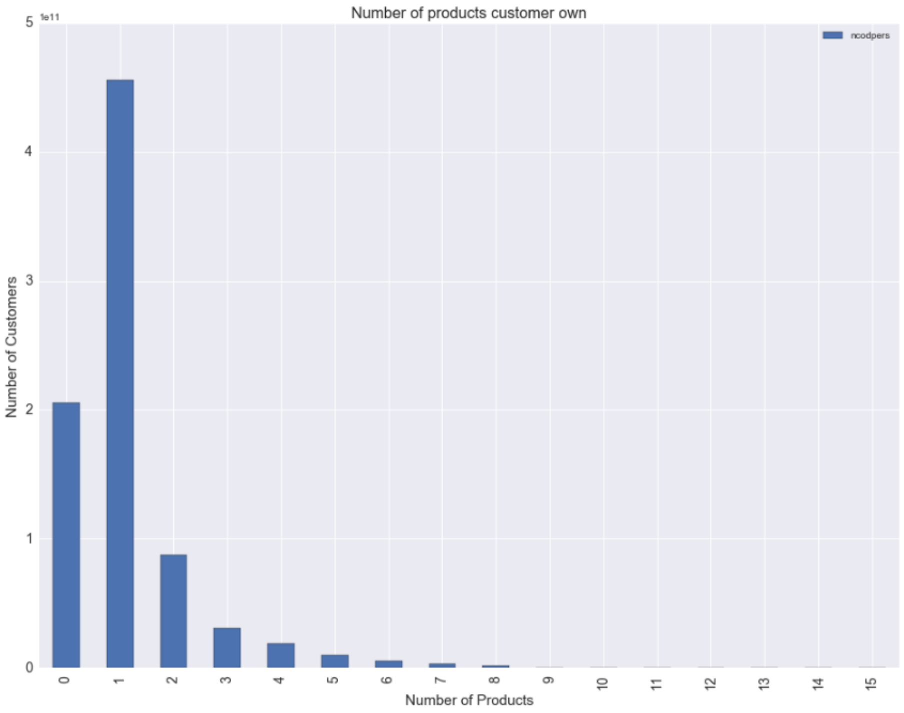
EDA

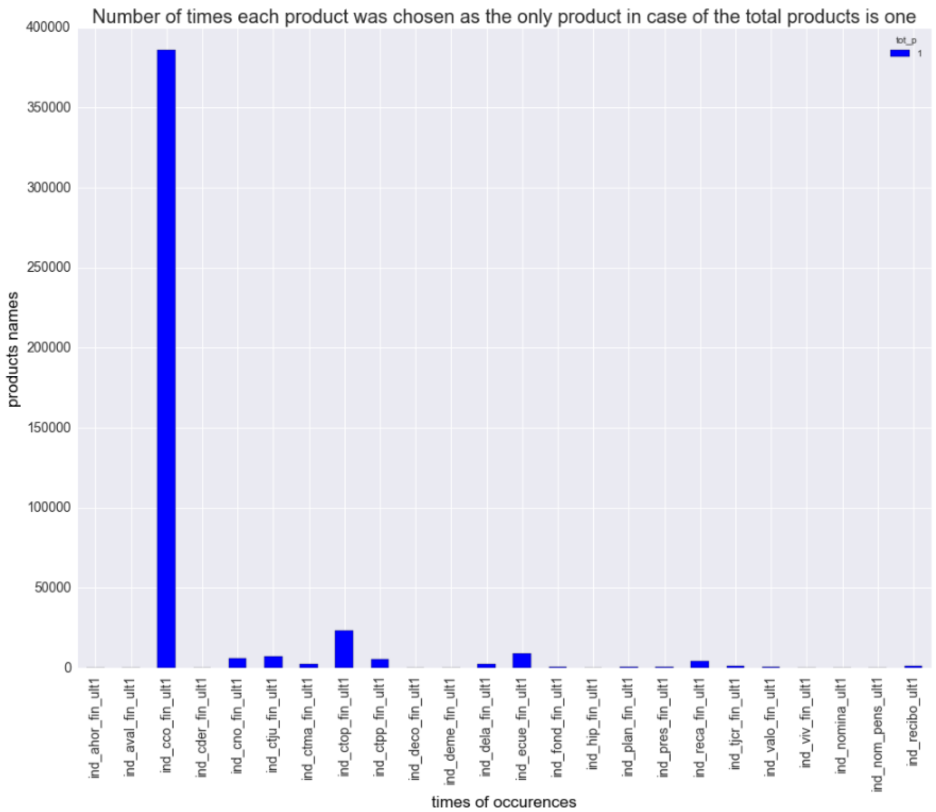
Number of Customers by Time



EDA

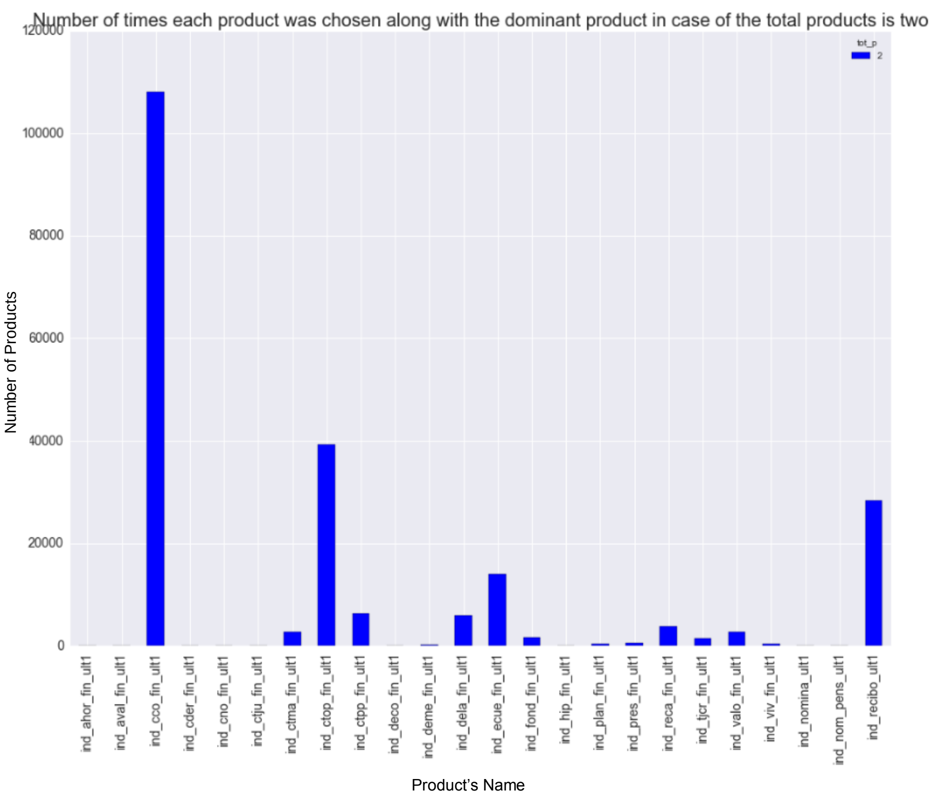
Number of Product Own - 2016.5

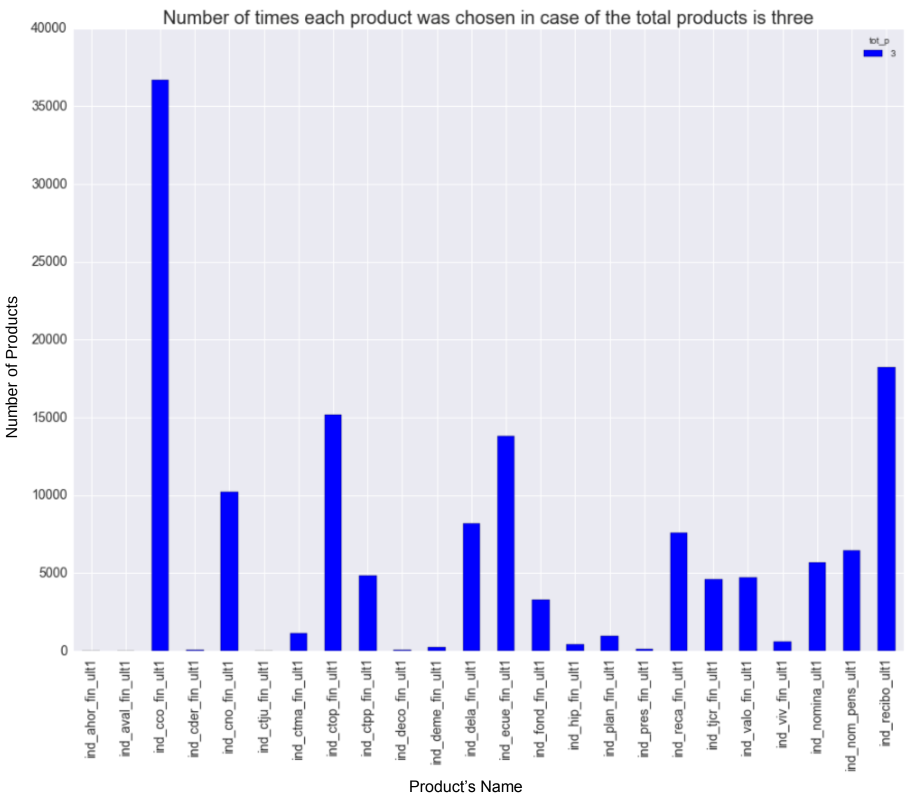




EDA

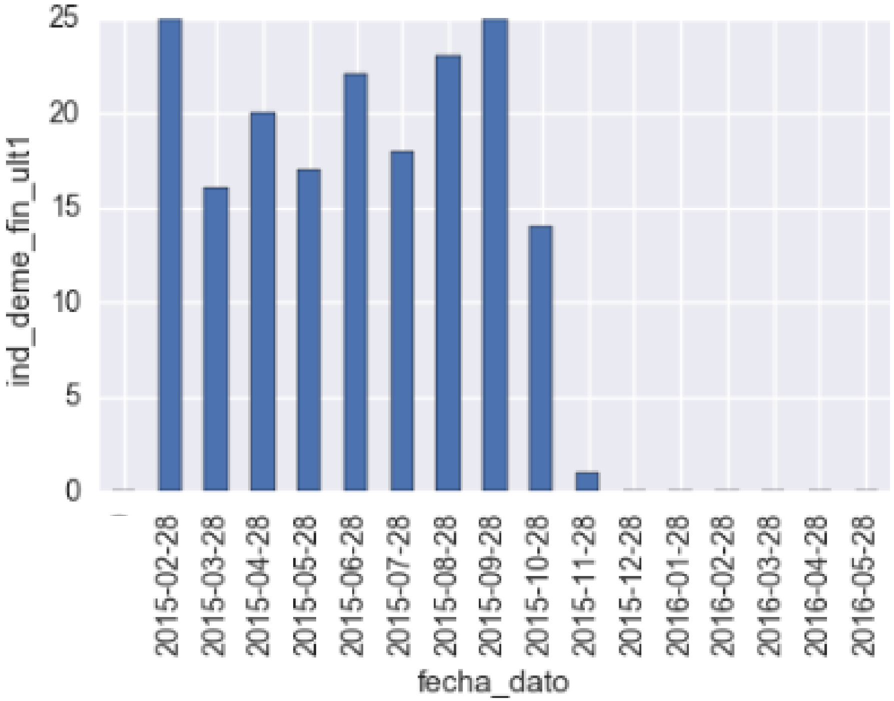
Number of Product Own - 2016.5

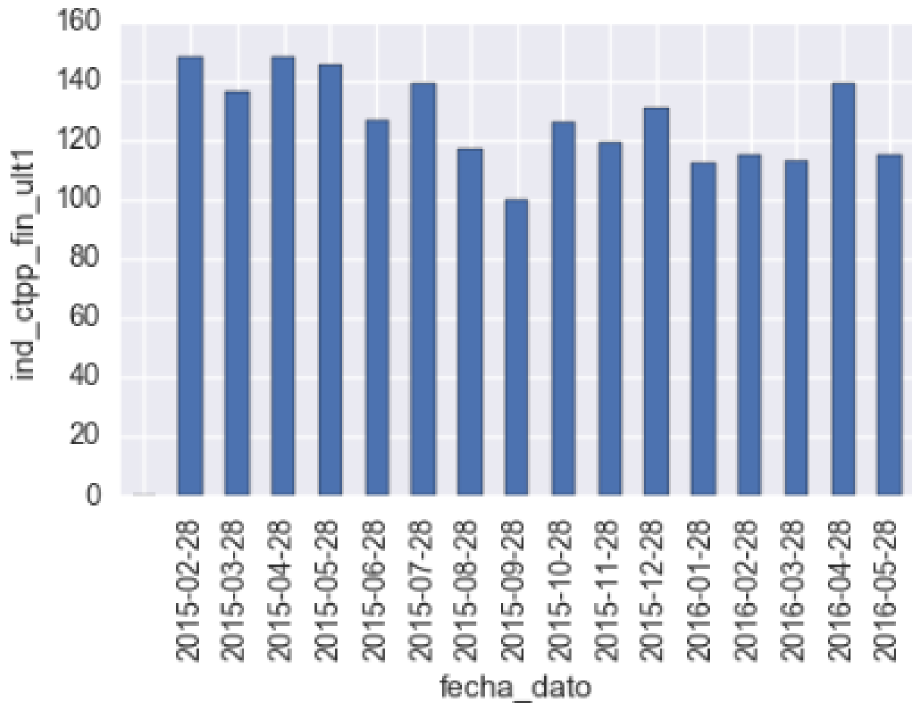




EDA

Number of Product Sales by Time





Feature Engineering

Input Features

Encoding

Output Features

Encoding

Use adjacent month
i.e. 2016.1-2016.2

Use the same month
i.e. 2015.5 – 2016.5

Use the seasonal month
i.e. 2016.3 – 2016.6

Input Features

Previous Month
Products

Use adjacent month
i.e. 2016.1-2016.2

Use the same month
i.e. 2015.5 – 2016.5

Use the seasonal month
i.e. 2016.3 – 2016.6

Input Features

Create Change
Features

i.e. Current -
Previous

Time Series
Pick significant pattern
Level = 0 , 1
&
Create as new
input features

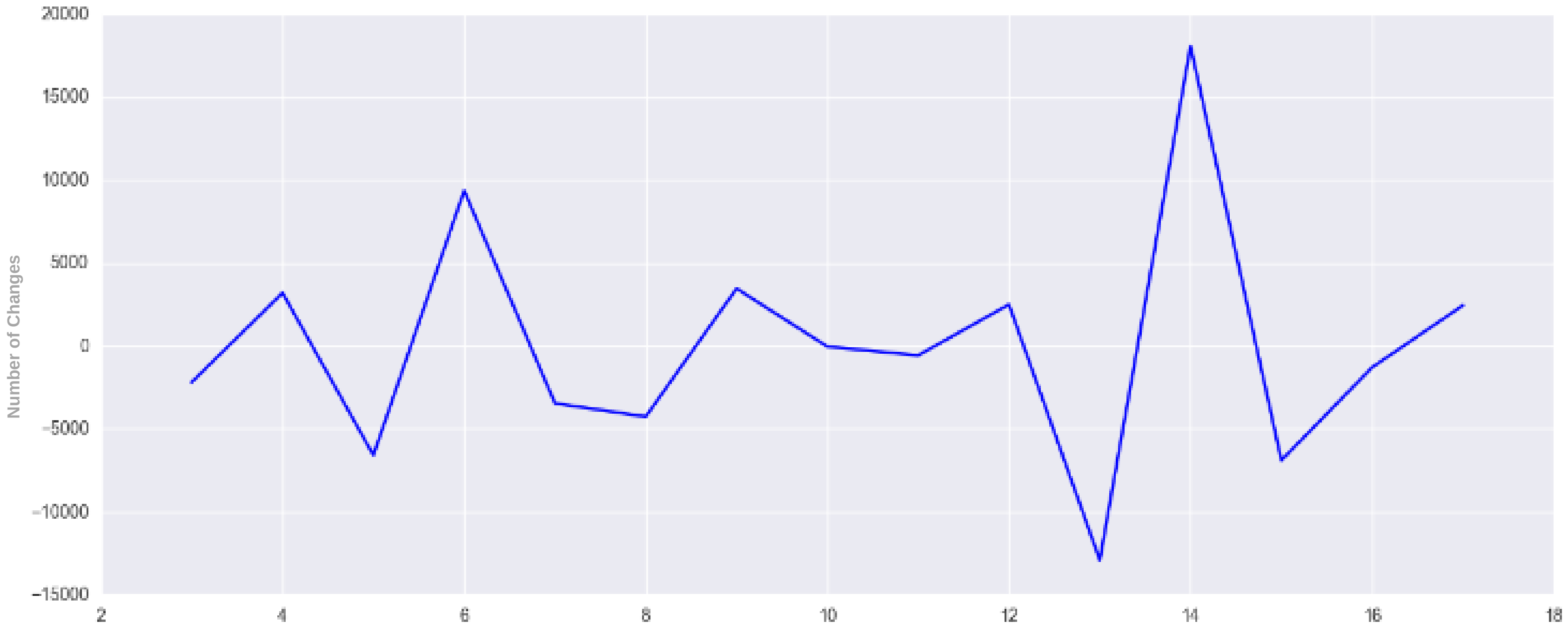
Input Features

Time Series
Level = -1, 0, 1

Output Features

Drop features & add weight
Based on popularity of the products

Time Series



Results of Dickey - Fuller Test

Pension Account

Test Statistic	-3.163039
p-value	0.022226
No. Lags Used	4.000000
Critical Value (5%)	-3.232950
Critical Value (1%)	-4.331573
Critical Value (10%)	-2.748700

Models Training

Random Forest	0.0295
Naïve Bayes & Genetic Algorithm	0.0272
XGBoost	0.02996
Neural Network	0.01995
AdaBoost	0.012
Collaborative Filtering	0.02364

Make recommendation based on products' popularity | 0.0225

Ensemble - Voting

01 Popularity

02 Collaborative Filter

03 Naïve Bayes

0.0276

01 Popularity

02 Collaborative Filter

03 Naïve Bayes

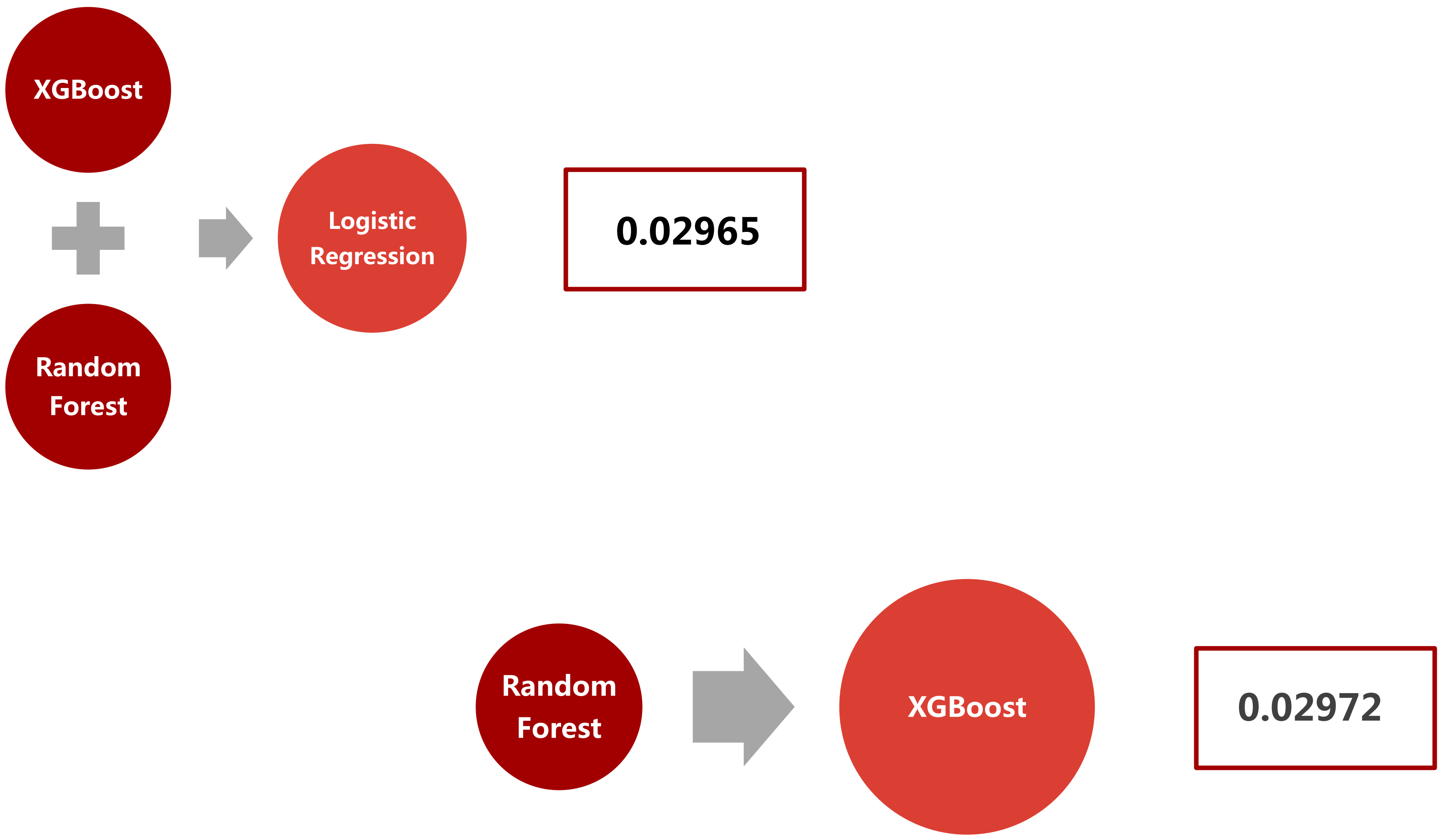
04 XGBoost

05 XGBoost (new Features)

06 Random Forest (multiclass)

0.0230108

Ensemble - Stacking



Insights & Findings

Using the month from the previous year has better prediction than the previous month from the current year

Single Model, XGBoost has the best performance

The best performance model contains features from five previous months

Multiclass has better performance than multi-labels

Final Result

173	↑72	TeraFlops	0.0299646	76	Tue, 20 Dec 2016 12:50:57 (-24.1h)
174	new	**Lydia Kan**	**0.0299626**	**10**	**Tue, 20 Dec 2016 14:47:15**
175	↑274	FJR2	0.0299618	26	Tue, 20 Dec 2016 15:56:57
176	↑258	Riju Bhattacharyya	0.0299613	37	Mon, 19 Dec 2016 14:34:26 (-18.5h)
177	↑525	三个和尚没水喝	0.0299611	38	Tue, 20 Dec 2016 06:40:13 (-31h)

Total Teams : 1806

Top 9 %