CSC 418

GROUP 1 DS PROJECT - **MAJOR**

CUSTOMER VALUE PREDICTION

GROUP MEMBERS

MBURU RYAN	-	P15 / 130645 / 2018
KAHURA GEORGE NDUNGU	-	P15 / 129315 / 2018
DIANGA MICHAEL OBADHA	-	P15 / 136909 / 2019
KANDIE JEFF	-	P15 / 137815 / 2019
HASSAN AHMED	-	P15 / 129439 / 2018

Github link —> https://github.com/mbururyan/Customer-Value-Prediction

CSC 418

GROUP 1 PROJECT - MAJOR

DATA SCIENCE

CUSTOMER VALUE PREDICTION

1. Business Understanding
    - Business Problem / Problem Statement
    - Business Solution
    - Business Background / Literature Review
        - Santander Bank
        - Customer Lifetime Value
        - Data Science in Banking
    - Resource Inventory
    - Project Plan
    - Methodology
2. Data Understanding
    - Source of data
    - Dimensions of the data
3. Data Preparation / Cleaning
    - Missing Data
    - Duplicate Data
    - Constant Columns
4. Data Exploration
    - Data Distribution
5. Feature Engineering
    - Data Splitting
    - Dimension reduction
6. Modeling (pre-PCA)
    - Linear Regression
    - LGBM
    - XGBoost
7. Modeling (post-PCA)
    - Linear Regression
    - Ridge Regression
    - Lasso Regression
    - SVM

~ USED THE CRISP-DM FORMAT TO DOCUMENT ~

1. Business Understanding

Business Problem / Problem Statement

Santander bank, wants the help of data scientists to assist them predict the customer value of their clients based on their transaction data, where the said data is anonymized for privacy reasons. This will assist their decision making in their business and marketing strategies.

Data science, machine learning techniques and business minded thinking is crucial to solving this problem and coming up with the best prediction model for evaluation of a customer's value.

Business Solution

Through Data Science techniques, we are meant to predict the value of each transaction by the Santander Bank customers. This will assist the bank in the decision making as earlier stated.

The prediction will be handled by Machine learning, where manipulated data will be fed onto different machine learning models then compare them. The best performing model will be picked and used by the bank's machine learning specialists hopefully.

Business Background / Literature Review

Santander Bank

Santander Bank is a multinational bank originating from Santander, Spain. It was founded in 1857 as a local savings bank. Over the years, the bank expanded both domestically and internationally, acquiring several other banks and financial institutions along the way.

In the late 20th and early 21st centuries, Santander became a major player in the global banking industry, with a presence in many countries across Europe, Latin America, and North America. Today, Santander is one of the largest banks in the world, offering a wide range of financial products and services to its customers.

Currently, the bank is owned by the Spanish government, and is a subsidiary of the Spanish Santander group. Its HQ being in Boston, it has $57.6 million in deposits and has over 3000 ATM's. The bank also employs close to 10,000 customers.

Customer Lifetime Value

Customer Lifetime Value (CLV) is the measure of the total income a business (in our case Santander Bank) can expect to generate / make from a particular customer as long as he/she continues to be a customer.

When the CLV value is being measured, often the best pointers to look at is the total revenue generated and the average profit made by the business / bank.

Why CLV is important to businesses :

1. **You can't improve what you don't measure** - Once the business starts measuring customer lifetime value and breaking down the various components, they can employ specific strategies around pricing, sales, advertising and customer retention with a goal of continuously reducing costs and increasing profit.

2. **Better Decision Making** - Once the business knows the value of a customer, they can make decisions on marketing to him/her so as to avoid losing them to competition etc.

3. **Improved Forecasting** - CLV forecasts help you make forward-looking decisions around inventory, staffing, production capacity and other costs. Without a forecast, you could unknowingly overspend and waste money or underspend and put yourself in a bind where you struggle to keep up with demand

4. **Improved sales** - Some retailers, tech companies, restaurant chains and other businesses have loyal customer bases that come back again and again. You can use CLV to track the average number of visits per year or over the customer lifetime and use that data to strategize ways to increase repeat business

5. **Increased Profitability** - Overall, a higher CLV should lead to bigger profits. By keeping customers longer and building a business that encourages them to spend more, you should see the benefit show up on your bottom line

Data Science in Banking

Data Science plays a crucial role in the banking industry. It helps banks to handle, analyze and manipulate large amounts of data to make informed business decisions, improve operational efficiency, and enhance the customer experience of their clients.

Some applications of Data Science in banking include:

1. **Customer behavior analysis** - banks use data to better understand their customers, including their spending patterns, financial goals, and risk profile.

2. **Fraud detection** - banks use data and machine learning algorithms to identify and prevent fraudulent activity.

3. **Risk management** - banks use data to assess and manage risks associated with loans, investments, and other financial products.

4. **Marketing and sales** - banks use data to target their marketing efforts and personalize their products and services to meet the specific needs of individual customers.

5. **Operations and efficiency** - banks use data to optimize their internal processes and reduce costs.

Overall, Data Science has become an essential tool for banks to stay competitive and meet the changing needs of the ever-evolving market.

Resource Inventory

Resources needed for this project include :

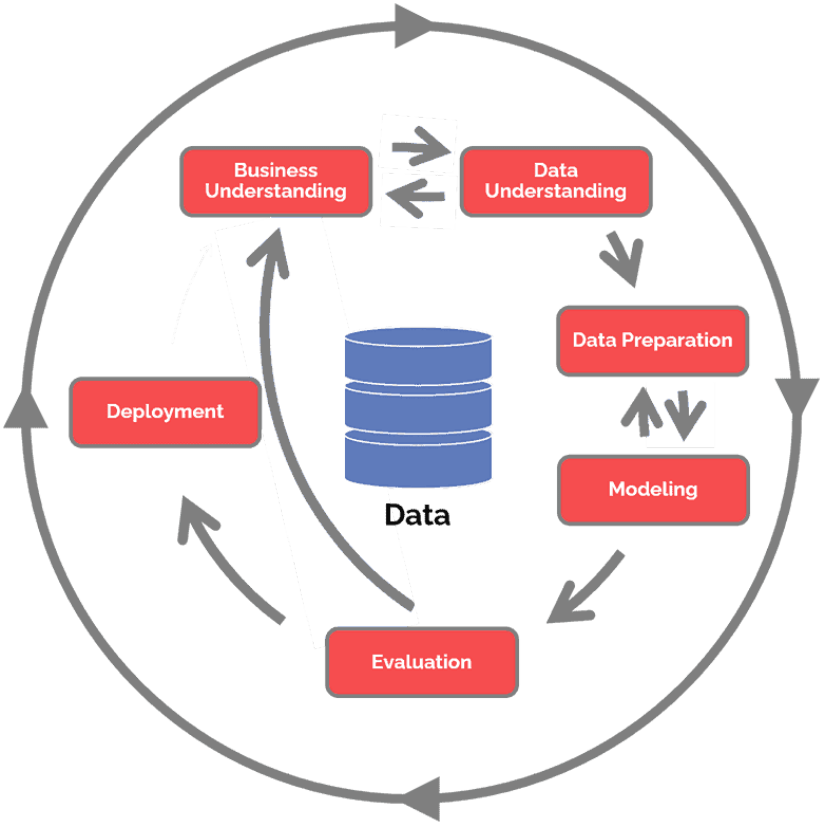
1. Datasets from the bank
2. Laptop with internet connectivity
3. Jupyter lab access
4. Data scientists (group members)

Project Plan

Phase	Time	Resources
Data Understanding	3 days	All Group Members
Data cleaning	5 days	All Group Members
EDA	3 days	All Group Members
Feature Engineering	1 week	All Group Members
Modeling	2 weeks	All Group Members
Documentation	1 week	All Group Members

The whole project was to be done in under a month, with collaboration from every group member.

Methodology



That is the Cross Industry Standard Process for data Mining methodology. It has 5-6 essential steps that have to be followed so as to thoroughly understand and manipulate the data at hand as intended :

1. **Business Understanding** - This is the current stage we are on, where the problem at hand is examined and researched on. Once a conclusive plan to how the problem will be solved is identified, a plan is laid out and resources outlined.

2. **Data Understanding**. This is where the data at hand is looked at from afar and the dimensions, size etc are outlined.

3. **Data Preparation**. This is where the data is checked for outliers, null values, duplicates and other inconsistencies and handled if any exists. EDA also happens here

4. **Modeling**. This is where the clean data is run through specified Machine Learning models.

5. **Evaluation**. This is where the performance of the ML models are evaluated and compared .

6. **Deployment**. If necessary, the models are deployed for real-time use by systems, websites, mobile applications etc.

2. Data Understanding

Source of data

The data was collected from Kaggle. The information was provided by our lecturers .

https://www.kaggle.com/competitions/santander-value-prediction-challenge

Dimensions of the data

The dataset has 4459 rows of data.

The dataset also has 4493 columns of data. Though most of the data is anonymized due to security reasons.

The dataset mostly consisted of numerical data.

3. Data Preparation / Cleaning

Missing Data

The data had 0 missing values.

Duplicate Data

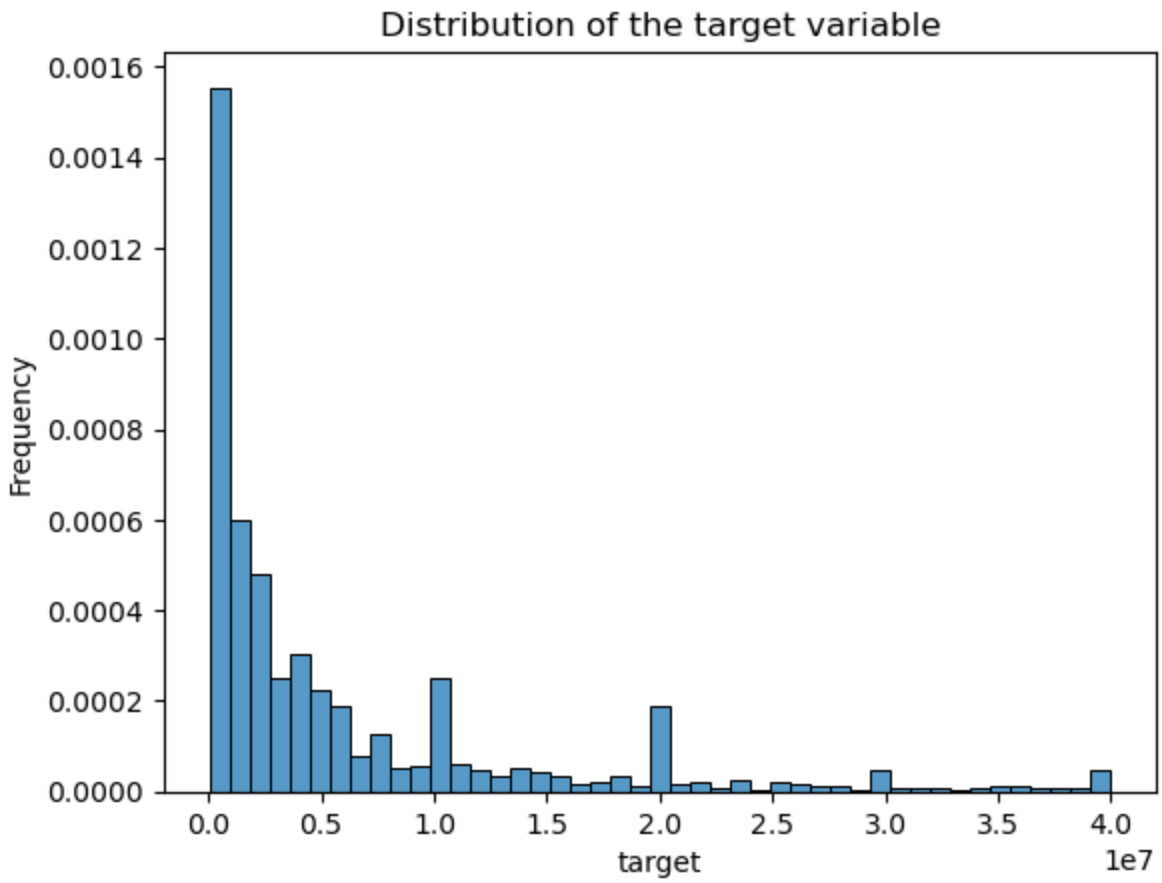
The data had 0 duplicate values.

Constant Columns

The data had 256 constant columns. The constant columns were instantly dropped.

4. Data Exploration

Data Distribution



As seen, the analysis that can be made is that transactions are on the 'lower end of the spectrum'. This is so as most of the data falls in the range of 0.0 - 0.5

5. Feature Engineering

This is where the data is prepared for modeling.

Data Splitting

The data was split into training data and validation data, in a 80 - 20% split.

Dimension reduction

Due to the number of columns, this was bound to cause problems when modeling. PCA was implemented in a separate notebook, so as to preserve the integrity of the bank's data.

6. Modeling (pre-PCA)

Linear Regression

RMSE : 4212660564080597.5

LGBM

RMSE : 1.457

XGBoost

RMSE : 1.418

	A	B	C	D	E	F
1	**Exp. No**	**Model Type**	**Date**	**Dataset**	**RMSE**	**Choice**
2	Exp 1	Linear Regression	22/12/2022	Train	42,660,564,081	
3	Exp 2	Linear Regression	22/12/2022	Test	40,669,778,225	
4	Exp 3	LGBM	23/12/2022	Test	1.475	
5	Exp 4	XGBoost	28/12/2022	Test	1.418	YES

7. Modeling (post-PCA)

Linear Regression

RMSE : 1.844

Ridge Regression

RMSE : 1.844

Lasso Regression

RMSE : 1.844

SVM

RMSE : 1.645

Decision Trees

RMSE : 2.288

Random Forests

RMSE : 1.626

Random Forests (with params)

RMSE : 1.57

LGBM

RMSE : 1.60

	A	B	C	D	E	F
1	**Exp. No**	**Model Type**	**Date**	**Dataset**	**RMSE**	**Choice**
2	Exp 5	Linear Regression	2/1/23	Train	1.844	
3	Exp 6	Ridge Regression	2/1/23	Train	1.844	
4	Exp 7	Lasso regression	2/1/23	Train	1.844	
5	Exp 8	SVM	3/1/23	Train	1.645	
6	Exp 9	Decision Trees	3/1/23	Train	2.288	
7	Exp 10	Random Forests	3/1/23	Train	1.626	
8	Exp 11	RF - (*after Hyper Param Tuning)*	5/1/23	Train	1.57	**YES**
9	Exp 12	LGBM	6/1/23	Train	1.6	

8. Conclusion (Findings)

This was an interesting project due to the anonymity of the data, and the team had to get creative when modeling. The data was fairly clean and not alot of data analysis was possible as most of the data was anonymized, so no interesting business insights could be extracted.

We dived deep into Machine Learning and after the conclusion of the project, the following models were to be extracted for deployment by the bank if necessary.

1. XGBoost model. This model had a performance of 1.4 when using root mean squared error. The original features were not reduced in this model, and we deem this the best one.
2. PCA model. If the clients prefer the dimension approach, external data has to be passed through the model to be reduced to 6 features so as to avoid inconsistency.
3. Random Forests. If the reduction route is chosen, the tweaked Random Forest model is the best one with an RMSE of 1.57

To preserve the integrity of the bank, we went with two approaches and will provide the 3 models as our final product. Thank you !