Santander Customer Satisfaction

A Project Report

Submitted in Partial Fulfillment of Requirements for Sharpest Minds Mentorship Program

Submitted By – Swaroop Todankar

Date – 24th October 2019

[1]

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

1. Introduction

Santander Bank is a North American bank owned by the Spanish Santander Group which employs around 10 thousand employees and has about 57 Billion dollars in deposits. It has its principal market focused around the north-eastern states of America [2].

For any consumer focused business there broadly exist two types of consumer mindsets: satisfied and dissatisfied. It also happens in very few cases that the dissatisfied consumers openly voice their opinion before leaving. This brings further to the question- How can the company improve upon the areas where the customers are dissatisfied? Such a question was put forth by Santander Bank in order to gain insights regarding the satisfaction level of customers.

The bank provided a large dataset consisting of about 76 thousand rows and 371 features. The expected outcome was a model capable of predicting dissatisfied customers early in the process so that necessary steps can be taken before it was pretty late.

2. Business Problem

The main question is- If the organization was to reduce the means by which customers are affected or left dissatisfied – What can be done for such a situation? Can early prediction of such a customer be possible? What measures can be taken to retain the customer?

This report aims to put forward an analysis of prediction of customer satisfaction using machine learning algorithms such as Random Forest and Gradient Boosting.

3. People interested in the Project – Target Audience

The target audience in this scenario are the management of Santander Group who are responsible to make the necessary decisions once the factors contributing to the dissatisfaction are identified.

The factors can be identified by paying extra attention to the customers identified by the algorithm and tending to their needs.

4. Data required

The data required to build a model to predict the customer satisfaction is as follows:

Customer Data: Data pertaining to the customers is provided by the bank for analysis.

5. Methodology

The following steps were employed to obtain the required results:

5.1 Importing Necessary Libraries

The first step is to import the necessary libraries and packages.

- Numpy – For numerical calculations
- Matplotlib – plotting and visualization
- Pandas – Data manipulation
- Sklearn – Machine Learning
- XGBoost – Gradient Boosting

5.2 Google Drive Pre- requisites

This step involves the authentication steps taken in order to use the dataset stored in Google Drive which can be directly loaded into Google Colab.

5.3 Pre-Processing

The following preprocessing steps were employed:

1. Checking the columns (features) present in the dataset

2. Getting statistical information of the dataset

3. Checking for presence of null values

4. Correcting outliers present in one feature (var3)

5. Checking is the dataset is balanced or unbalanced

6. Performing the train- test split

7. Scaling the training splits

5.4 Machine Learning

The following 2 machine learning algorithms were used:

1. Random Forest Classifier

2. Gradient Boosting Classifier

5.4 Analysis

The analysis involved checking the precision and recall score of above machine learning algorithms in two sections which are as follows:

1. Section A – Unbalanced dataset

This section involved using the dataset after preprocessing without performing any sampling operation. The two machine learning models were trained and were evaluated and the metrics were calculated.

2. Section B – Balanced dataset

In this section undersampling procedure was performed to balance the dataset (3008 rows). Further, using this balanced dataset the two machine learning models were trained and metrics for evaluation were obtained.

**** The models were trained on the balanced dataset and then tested on the unbalanced dataset to gauge the working power of the trained model.**

The models were subjected to Randomized Search Cross Validation and the best hyper parameters were obtained.

These parameters were used to train the models again and the results were fed into a dataframe.

5.5 Metrics for Evaluation

The following metrics were used for evaluation:

1. ROC- AUC

2. Precision

3. Recall

4. Log-loss

5. Confusion Matrix

6. Results and Discussion

6.1. Random Forest Classifier

	Accuracy Score	Log Loss	ROC AUC	Precision		Recall	
				[0]	[1]	[0]	[1]
Random Forest	95.34	0.73	0.670	0.96	0.14	0.99	0.03
Balanced Dataset	72.17	0.621	0.7920	0.74	0.72	0.73	0.73
Balanced Optimized	71.67	0.616	0.7945	0.74	0.71	0.72	0.74
Imbalanced Dataset	79.40	0.581	0.685	0.99	0.06	0.38	0.95
Imbalanced Optimized	87.338	0.561	0.6562	0.99	0.06	0.40	0.95

Inferences-

1. Using Random Forest Classifier model on the entire dataset with 25 features, the accuracy obtained was 95.34 % owing to the fact that the dataset was highly unbalanced (97- 3).

2. After balancing the dataset (Undersampling) with 3008 records and 25 features the accuracy decreases to 72.17 % (24.3% decrease) and log loss decreases to 0.621 (15% decrease) also AUC increases by a bit (18.2 % increase).

3. Using Randomized Search CV for hyperparameter optimization the score decreases by a bit (6% decrease) with a little decrease in log loss (8% decrease). AUC increases insignificantly (3% increase).The selected parameters were n_estimators = 400 and max_features =sqrt.

4. Using the trained model on balanced dataset, to predict the results on imbalanced dataset, the accuracy increases (10% increase) and loss further reduces (5% decrease). The AUC is lower in this step as compared to the previous ones (13 % decrease).

5. Optimizing the imbalanced dataset similar to the balanced dataset provides an accuracy of around 87 % (9% increase) with a loss of about 0.561(3% decrease) with further little decrease in AUC (4 % decrease)

6.2. Gradient Boosting Classifier

	Accuracy Score	Log Loss	ROC AUC	Precision		Recall	
				[0]	[1]	[0]	[1]
Gradient Boosting	96.00	0.13	0.838	0.96	0.00	1.00	0.00
Balanced Dataset	74.00	0.5139	0.8233	0.78	0.73	0.73	0.78
Balanced Optimized	74.833	0.516	0.8207	0.77	0.71	0.71	0.77
Imbalanced Dataset	63.752	0.639	0.6553	0.99	0.05	0.19	0.97
Imbalanced Optimized	90.706	0.4375	0.6449	1.00	0.06	0.37	0.96

Inferences-

1. Using Gradient Boosting Classifier model on the entire dataset with 25 features, the accuracy obtained was 96.00 % owing to the fact that the dataset was highly unbalanced (97- 3).

2. After balancing the dataset (Undersampling) with 3008 records and 25 features the accuracy decreases to 74 % (23% decrease) and log loss increases to 0.5139 (255% increase) also AUC decreases by a bit (1 % decrease).

3. Using Randomized Search CV for hyperparameter optimization the score increases by a bit (1% increase) with little change in log loss (0.4% increase) and AUC (0.3% decrease). The selected parameters were max_child_wt = 7, learning_rate = 0.05, max_depth = 8, gamma = 0, col_sample_bytree = 0.4

4. Using the trained model on balanced dataset, to predict the results on imbalanced dataset, the accuracy decreases to about 63.752 % (14% decrease) and loss further increases to 0.639 (23% increase). The AUC is lower in this step as compared to the previous ones (20% decrease).

5. Optimizing the imbalanced dataset similar to the balanced dataset provides best accuracy of around 90.706 % (42% increase) with a decrease in loss (31% decrease) with further decrease in AUC (1% decrease).

7. Limitations

The following inferences can be more refined if the names of the features were available. This would help in removal of unnecessary features and would further provide more idea into feature importance.

8. Conclusion

Using Google Collab, the dataset stored in google drive was loaded. Pre-processing steps were carried out and machine learning algorithms were applied. Hyper parameter tuning was carried out to obtain optimum results. The metrics for models were evaluated and compared.

9. References

1. Pymnts, “Santander Digital Investment Unit Global Banking Spanish Banks,” [Online]. Available: https://www.pymnts.com/news/international/2018/santander-digital-investment-unit-global-banking-spanish-banks
2. Wikipedia, “Santander Bank,” [Online]. Available: https://en.wikipedia.org/wiki/Santander_Bank