Banco Santander Customer Transaction Prediction.

Problem Statement:

Customer preservation is essential in a variety of businesses as acquiring new customers is often costlier than keeping the current ones. Many companies are therefore always trying to answer the question “How can we predict the value of a customer over the course of his or her interactions with the business”. Santander therefore wants to know future customer transaction predictions.

Company overview

Banco Santander, S.A. is a Spanish multinational commercial bank and financial services company founded and based in Santander, Spain. Their mission is to help promote businesses and people by helping them understand their financial status and how best to achieve their monetary goals. Santander is presently working on how to accelerate its digital transformation and platform strategy to boost growth and increase profitability.

One of these transformations is to identify which customers will make at least one single transaction in the future irrespective of the amount.

This question can best be answered by using predictive analysis of the customer past interactive data with the company and then a machine learning model is used to predict the possibility of these customers being retained or lost.

From these predictions, the company will know how much more efforts they need to keep the old customers which could be through bonuses and discounts, special coupons, better customer services and many more. While also thinking of how to attract more customers into the business through marketing campaigns, offer more discounts and deals, online adverts etc.

Source of Data and description

The data for this project is taken from Kaggle, an online community of Data Scientists and machine learners owned by Google LLC.

The data set contains two comma separated value files for training and testing on the model with anonymous feature names for security purposes.

The train data set contains:

- Unique ID_code
- 200 numerical features labeled from var_0 to var_199
- Target either a 0 for no future transaction and a 1 for future transaction
- Total of 20000 unique records

Methodology

This project will undergo a series of processes outlined as:

- Exploratory Data Analysis where the different relationships between the 200 features and the target
- Inferential Statistics applied on the different features and the target.
- After these analyses, stories will then be brought out from these results depending on the visualizations gotten.
- Since this project is a classification problem with two classes (0 and 1), different machine learning models will be tested on the dataset to see which best predicts with highest model accuracy score.
- Applicable extensions of project and it's impact to the business and/or customers too.

Data wrangling and Storytelling from the Santander data set

The data set provided for Banco Santander Customer Prediction doesn't need cleaning with all records with unique data information.

Doing analysis on both the trained and test data frames, so much information was extracted from it which brings out so much insights from the data.

From the description of train and test data frames, the mean, min, and max are slightly different whereas the standard deviations of both are quite close but large. Standard Deviation being a measure that is used to quantify the amount of variation. This implies a low Standard Deviation value indicates that data points turn to be close to mean while a high Standard Deviation value shows data points are spread out over a wide range of values.

Visualizing the relationship between each feature and the target using density plots of all the 200 features shows sum of all probabilities with the peak of the density plot showing the highest probability of the data points. Most of these features have very low peak values which shows they have little or no influence on a customer making a transaction in the future. Also, their low standard deviation values also show the wide range of variability, so measures could be brought upon these anonymous features to make the values within a shorter range and eliminate lower value data points too. This will therefore decrease the standard deviation and an increased probabilistic value influence to the target variable 1.

Since the aim of this project is to predict customers that will make a transaction in the future, like 5% of the features have the peak value at a probabilistic level of 0.6 which is the highest i.e. they have lower values of Standard Deviation. These few features though unidentified for security purposes could therefore be used as the main target on how to best keep these customers.

To conclude, this analysis shows most of the features instead have high probabilistic values on a customer not doing a transaction in the future. Since this is not the aim of the project, those with lower Standard Deviation values and with higher probability values greater then 0.5 could therefore be used further though the features are unidentified.

Inferential Statistics

Inferential statistics, is trying to reach decisions that extend beyond the immediate data alone. Inferential statistics is therefore used to make judgments and predictions that an observed variation or correlation between groups of features is a dependable one or not.
Correlation shows the extent to which the variables being compared fluctuate together.

Inferential statistics on the Santander data was concentrated on some of the features which had high probability values when density plots were plotted against the 200 features. Three of these features were considered and each of them, a Pearson Correlation was computed to see their relationship with the target.

Considering a Pearson Correlation between these features and the target.

Also, the Null Hypothesis is set for all the three features as: there is no correlation between each feature and the target variable.

The following observations were made;

- An r value calculated by Pearson Correlation inbuilt function as 0.04 which is less than 0.02 implies there is a positive but weak correlation between the two features. This means the increase in the unidentified feature increases the probability of a customer making a future transaction.
- Also, the p value displayed is greater than 0.05 indicates a weak evidence against the null hypothesis. This therefore implies a fail to reject the null hypothesis. Moreover, we can't say if the correlation is statistically significant or not.
- For var_108 and var_148, the have a weak negative correlation with the target. Since their r values are all negative. This implies a decrease in the value of either of these features increases the probability of a client making a future transaction.

From this analysis, the Pearson correlation therefore makes the var_91 feature statistically important as this feature could be used to increase future customer transactions which is the aim of this project. If for instance this feature was a coupon or some bonuses, increasing these offers if increased could therefore increased the probability of having a customer make a future transaction.

Applying Machine Learning Algorithms

Two models will be used to make predictions so as to choose the one with best accuracy score as the most suitable one. The models considered are;

1. Logistic Regression with and without regularization
2. Gaussian Naïve Bayes model

Logistic Regression without regularization

This model after being trained and then tested, produced accuracy score of 0.9148 for the training set and 0.9112 for the test set. From these results, the following conclusions were made:

- The model is suitable for this problem as both accuracy scores are greater than 0.9
- Secondly, there is no model overfitting as the differences between the training accuracy and the test accuracy is minimal.
- Visualizing the training and test accuracies after 10 times of modeling the data, the accuracy range for training set was +/- 0.004 and that for the test set was +/-0.006 which is still too small to cause model overfitting.
- Classification report when testing the model
    - Under precision, this shows the ability of the classifier not to register an instance positive when it is actually negative. With a 0.93 probability to predict a 0 which implies a customer won't make a future transaction and a 0.62 probability to predict a 1 which is a customer will make a future transaction. This probability of predicting a 1 which is a future transaction is smaller because its occurrence is minimal in the dataset.
    - Recall, is the ability of the classifier to find all positive instances for each class it is defined. 0.98 probability of finding a 0 and 0.27 probability of finding a 1.
    - F1-score, is the weighted harmonic mean of precision and recall such that the best score is 1.0 and the worst is 0.0.This score is 0.95 for a 0 and 0.37 for a 1.
    - Support, is the number of actual occurrence of the class in the dataset. 4511 for a 0 and 489 for a 1.

Logistic Regression with Regularization.

From above observations, there is no model overfitting. Now the question:

Why use regularization?

Since the goal of every data scientist is to have accurate model performance as much as possible, regularization is therefore implemented to better model performance. Improvements of this model are seen below when analyzing the classifiers performance.

Accuracy score when training the model with regularization is 0.9124 which is slightly greater than that without regularization by 0.0009. This difference may seem minimal but it will greatly increase the probability of a customer making a future transaction.

Gaussian Naïve Bayes Model

Training this model produced an accuracy value of 0.9262 and when testing the model, it produced an accuracy value of 0.8606. This great difference between the train accuracy and the test accuracy therefore shows there is model overfitting.

Model overfitting is a situation where the model is too complex to explain features in the data. Also, overfitting causes a modelling error which occurs when a function is too closely fit to a limited set of data points.

Also, this model is not suitable as the test accuracy is less than 0.9 which is always the minimum value any suitable model should have for better predictions.

Classification report for Gaussian Naïve Bayes results shows more details after training and then testing the model.

Considering the test classification model:

- Under precision, this shows the ability of the classifier not to register an instance positive when it is actually negative. A 0.90 probability to predict a 0 and a 0.11 probability to predict a 1 which is a customer will make a future transaction.
- Recall, which is the ability of the classifier to find all positive instances for each class it is defined.0.95 probability of finding a 0 and 0.06 probability of finding a 1 which is really low.
- F1-score, is the weighted harmonic mean of precision and recall such that the best score is 1.0 and the worst is 0.0. A score of 0.92 for a 0 and 0.08 for a 1.
- Support, is the number of actual occurrence of the class in the dataset. 4511 for a 0 and 489 for a 1.

From these observations, Logistic Regression with Regularization is therefore the most suitable model to explain features in the data. Since this model had better performance and higher accuracy values from Logistic Regression without regularization, it is therefore the most suitable model compared with the other two models.

Comparing both Machine learning Algorithms implemented.

- Applying Logistic Regression on the data set produced good accuracy scores for both training and test data. But when displaying a classification report, the precision values for both classes has a large difference as the data is imbalance.
- For Gaussian Naïve Bayes, the model is overfitting because the training and test accuracies have a large difference which implies the model is too complex for the dataset.
- Therefore, the best algorithm for this case is Logistic Regression which is implemented after the data has been made balanced. This therefore solves the problem of Imbalance data which was faced because of the great difference in the two classes.

Further Research

- Further research has to be done on how to best solve the issue of imbalance data as the two target classes have a great difference. This can therefore be solved by:
- **Resampling the Dataset:** This is done by randomly adding copies of instances from the under-represented class called over-sampling. This is most suitable for the minority class which therefore increases its number of samples.
- **Generating Synthetic Samples:** SMOTE (Synthetic Minority Oversampling TEchnique) system works by creating synthetic samples from the minor class instead of creating copies. The algorithm selects two or more similar instances (using a distance measure) and perturbing an instance one attribute at a time by a random amount within the difference to the neighboring instances. These measures put in place will reduce the problem of data unbalancing.
- After implementing the techniques above on the imbalanced data, Logistic Regression can therefore be used to better do training and testing of the model which will result in better predictions and accuracies as the two target classes are already made balanced.

Client Recommendations

- Considering the three features with highest probability values of having a customer make a future transaction in the future. These features are var_91, var_108 and var_148 produce the highest probability values of having a 1 as target. These features though anonymous could therefore be used to further increase this probability by ensuring values equivalent to their mean or higher.
- Also, the features with lowest probability values like var_5, var_84 and var_86 are the one of the features causing the customers not to make a future transaction in the future. These features which are unidentified could further be used by ensuring they instead contribute to a higher probability of having a 1 as the target.
- From this analysis, the average number of customers transaction per day, week or month could be calculated and from this the company will know if they are meeting up with the target ahead of them or not. And if the target is not being met, they can therefore bring forth improvements to this by offering bonuses and more promotions which will increase customer awareness about their products and services.

References

- ***Think Stats: Probability and Statistics for Programmers by Allen B. Downey***
- ***Significance test (Hypothesis testing) on Khan Academy***
- ***Logistic Regression Analysis by CM Dayton***
- ***Introduction to Machine learning by Alex Smola***