Customers Churn Predictions

Raju Ranjan
rajur.ug18.ec@nitp.ac.in
B.Tech., Department of
Electronics and Communication Engineering
National Institute of Technology Patna, India

Shruti Sinha
shrutis.ug18.ec@nitp.ac.in
B.Tech., Department of
Electronics and Communication Engineering
National Institute of Technology Patna, India

Abstract

Customer churn is a major problem and one of the most important concerns for large companies and banks. Due to the direct effect on the revenues of the companies, especially to take necessary actions to reduce this churn. The main contribution of our work is to develop a churn prediction model which assists Banks to predict customers who are most likely subject to churn. The model developed in this work uses machine learning techniques on big data platform and builds a new way of features engineering and selection. In order to measure the performance of the model, the Area Under Curve (AUC) standard measure is adopted.

Keywords -

Customer churn prediction, Machine learning, Feature engineering, Logistic regression, AUC - ROC, Precision, Recall

1 Introduction

During the enormous increase in numbers of customers who are using the communication sector and in numbers of companies the competitive level between companies raised. Each company tries to survive in this competition through many strategies, The Main strategies are: 1) upsell existing customer,increase duration of retention of their customers, acquire new customers. Companies are concerned about seeking to keep or retain their customers as they are considered that as a profit, and it is cheaper to keep them than to earn a new one. Each company tries to keep its customers, by make them more loyal. Customers are great ambassadors in the market [5] as the company can use them for making advertising of the company's product or service.

A customer churn happens when customers are not satisfied with a service provided by a company. It results in customers switching to another service provider. Customers have different reasons for churn, and all of them should not be treated in the same way. There is a need for a prediction model to predict churn customers and provide a strategy of retention depends on their churn factors.

In this paper we supposed a situation in which let say Banks wants you to predict whether a bank customer will churn or not or whether the customer will be able to maintain the minimum required average monthly balance or not. This model is designed using classification algorithm based on logistic regression.

2 Data Dictionary

There are multiple variables in the dataset which can be cleanly divided in 3 categories:

1. Demographic information about customers
    - customer_id : - Customer id
    - vintage :- Vintage of the customer with the bank in number of days
    - age :- Age of customer
    - gender :- Gender of customer
    - dependents :- Number of dependents
    - occupation :- Occupation of the customer
    - city :- City of customer (anonymised)
2. Customer Bank Relationship
    - customer_nw_category :- Net worth of customer (3:Low 2:Medium 1:High)
    - branch_code :- Branch Code for customer account
    - days_since_last_transaction - No of Days Since Last Credit in Last 1 year
3. Transactional Information
    - current_balance :- Balance as of today
    - previous_month_end_balance :- End of Month Balance of previous month
    - average_monthly_balance_prevQ :- Average monthly balances (AMB) in Previous Quarter
    - average_monthly_balance_prevQ2 :- Average monthly balances (AMB) in previous to previous quarter
    - current_month_credit :- Total Credit Amount current month

- previous_month_credit :- Total Credit Amount previous month
- current_month_debit :- Total Debit Amount current month
- previous_month_debit :- Total Debit Amount previous month
- current_month_balance :- Average Balance of current month
- previous_month_balance :- Average Balance of previous month
- churn :- Average balance of customer falls below minimum balance in the next quarter (1/0)

.

3 Methodology

3.1 Classification Algorithm

In the Classification algorithm, a Categorical Target variable is used. We can work with Probability.We will make the predictions as probability between 0 and 1

3.2 Data Exploration and Preprocessing

3.2.1 Target Variable

The questions one should usually find answers to while exploring the target variable

- What information does this variable represent?
- What is the data type for this variable?
- Do the few sample values that we can 'eyeball' (i.e. skim through) make sense?
- What are the minimum, maximum, mean, median values for this variable and whether these make sense and what we can infer from them?
- Does the target variable contain any Outliers that we need to treat?
- Are there any missing values in the target variable that we need to treat?
- What is the distribution of the values of this variable over its range? Are the values uniformly or normally distributed or is data skewed towards lower values (i.e. contains more of lower values) or towards higher values?

3.2.2 Categorical Variables

- Find the frequency distribution of unique values of a categorical variable.
- Visualize the relationship between a numerical target variable and a categorical variable with the help of a bar chart.
- The questions one needs to answer while exploring categorical variables
  - What information does this variable contain
  - Do the few sample values that we are able 'eyeball'
  - How many and what unique values , i.e, level, does this categorical variable have and what is the frequency distribution of those unique values
  - How is dependent/target variable correlated with this variable.
- The method of finding out whether a Categorical variable is having an impact on the numeric variable is known as **ANOVA**
- **ANOVA**- it checks if the means of the target variables across different levels or unique values of a categorical variable are equal or not.
- **Binning and Creation of dummy variables:**
  - Transforming a categorical variable into a set of numerical or boolean variables called dummy variables, each of which has values of 0 or 1.
  - If Categorical variables has 'n' levels $\rightarrow$ n-1 dummies are required

  Why we are doing this :
  - Regression modeling requires all the independent variables to be numerical variables Get_dummies of pandas library creates the dummy variable and drop the categorical variable. Binning - if number of levels > 20 , then bin these levels into fewer groups before creating the dummy variables.
  - Used for both numerical and categorical variables.

3.3 Feature Scaling

Feature Scaling is a must for those algorithms where some measure of distance between data points is involved like Logistic Regression, Linear Regression, K Nearest Neighbors, Principal Component Analysis, etc.

Different Techniques of Feature Scaling:

- Standardisation : It rescales the feature values so that they have the properties of a Standard Normal Distribution with mean as 0 and the standard deviation of 1.
- Min Max Scaling: One of the simplest methods for scaling.The value range for the transformed variables lies between [0, 1]
- Normalization: The range is fixed from -1 to 1. Also called mean normalization
- The preprocessing library of sklearn can be used to do the standardisation.
- Standardisation method should be used for scaling the feature variables when a linear regression model is built.
- A linear regression model assumes the input variables to be normally distributed.
-

3.4 Feature Engineering

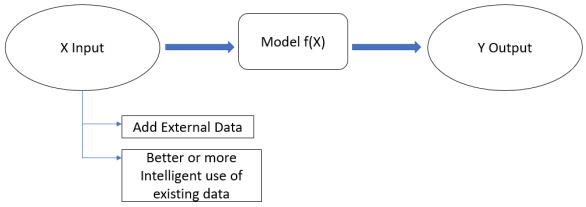

Figure 1. Flowchart: Feature Engineering

we conducted set of feature engineering step such as:

- The technique of generating new features using the existing features is called feature engineering.
- The science of extracting more information from existing data

We have the Preprocessed dataset names as churn_prediction.

3.5 Logistic Regression

- Logistic Regression is used when the dependent variable(target) is categorical.
- Logistic regression is basically a supervised classification algorithm. In a classification problem, the target variable(or output), y, can take only discrete values for given set of features(or inputs), X.

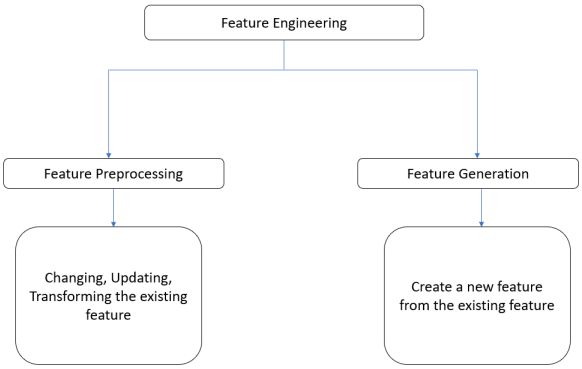

Figure 2. Types of Feature Engineering

$$S(x) = \frac{1}{1 + e^{-x}}$$

Figure 3. logit / sigmoid function

- Logistic regression uses logit Function also named as sigmoid function.A function which has a range between 0 and 1 (0 and 1 exclusive), irrespective of any input, it will always give output between 0 and 1.

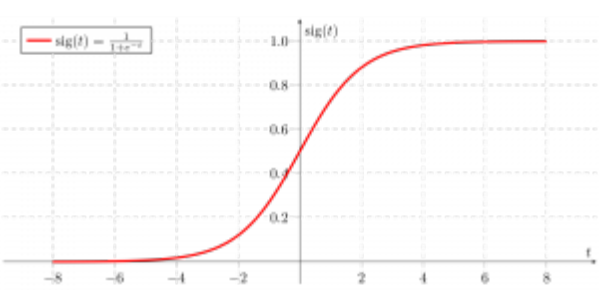

Figure 4. logit/ sigmoid function curve

- Predicting continuous probability values between 0 and 1 - Regression
- Gradient Descent can only work with the convex functions to find the ideal value of parameters.

figure

Overview In this implementation, we will be looking at the following steps: To predict this, we have the following information:

- Details of the customer such as age, gender, occupation, etc
- Details of the customer bank relationship such as customer_category, transaction_details, etc.

- Details of the customer bank relationship such as customer_category, transaction_details, etc
- Customer Account Balance details such as monthly average of previous months, current balance, etc.

1. Importing Libraries and Dataset :
   Python libraries like :
   - Numpy
   - Pandas
   - Matplotlib
   - Sklearn
2. checking the dataset distribution
3. separating dependent and independent variables
4. Scaling the dataset
   - The preprocessing library of sklearn can be used to do the standardisation.
5. splitting the dataset Data-
   - Training Dataset : To educate or train our models
   - Test Dataset : To examine model performance

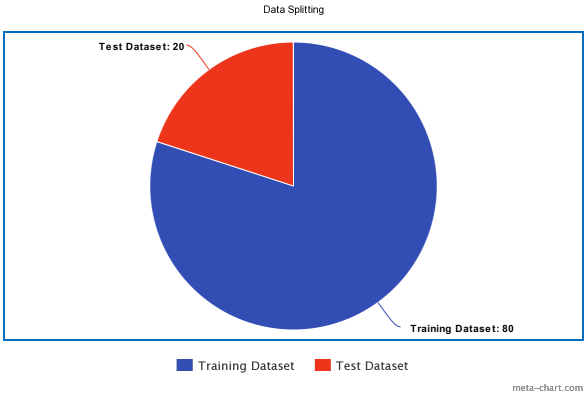

Figure 5. Pie chart for splitting of Dataset

6. Model Building, predictions and odds ratio
7. calculating the precision score
8. calculating recall score.
9. manually calculating the f1 score
10. gathering Precision/recall scores for different thresholds
11. AUC-ROC Curve
12. Coefficient Plot

$$F1 = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{recall}}}$$

Figure 6. F1 Score formula

3.6 Model Evaluation Metrices

To evaluate and choose the right model for classification problems.

- Confusion matrix:
  - It is used to interpret the model predictions systematically.
  - An n x n matrix, where 'n' represents the number of distinct classes in the target variable.
- The metrics which can be derived from the confusion matrix:
  1. Accuracy
  2. Precision
  3. Recall
- Accuracy : The higher the accuracy better the model

$$\text{Accuracy} = \frac{\text{Current prediction}}{\text{Total Prediction}}$$

Figure 7. Accuracy

- Using confusion matrix the formula become: .

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FN+FP}$$

Figure 8. Accuracy

- Imbalanced data : A dataset which has disproportionate numbers of either positive or negative classes and both the classes are not equally or nearly equally distributed
  Precision and recall handle the imbalanced dataset efficiently

- Precision : It is used when avoiding false positives is more essential than encountering false negatives.

$$\text{Precision} = \frac{\text{Predictions actually positive}}{\text{Total Prediction Positive}}$$

Figure 9. Precision

- Using confusion matrix the formula become:

$$\text{Precision} = \frac{TP}{TP+FP}$$

Figure 10. Precision

- Recall : It is used when avoiding false negatives is prioritized over encountering false positives.

$$\text{Recall} = \frac{\text{Predictions actually positive}}{\text{Total Actual Positive}}$$

Figure 11. Recall

- Using confusion matrix the formula become:

$$\text{Recall} = \frac{TP}{TP+FN}$$

Figure 12. Recall using confusion matrix

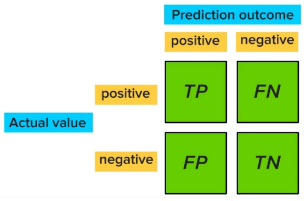

Figure 13. Prediction

- Log Loss
  - The cost function of the logistic regression.
  - Calculates the error of the classification Model
  - Farther a predicted probability is from its true class, higher is the log loss function
  - The log loss itself can help us distinguish between the apparent identical models.

$$H_p(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot log(p(y_i)) + (1-y_i) \cdot log(1-p(y_i))$$

Binary Cross-Entropy / Log Loss

Figure 14. Logloss formula

- AUC- ROC
  - AUC stands for Area under Curve
  - ROC stands for Reciever operating characteristics
  - AUC - ROC curve is a performance measurement for classification problems at various thresholds settings
  - Higher the AUC, the better the model is distinguishing between 0 and 1
  - If the AUC-ROC is >0.95, there could be something wrong with the model or the dataset.

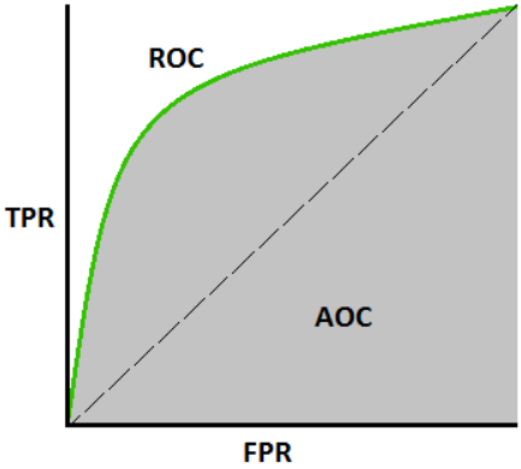

Figure 15. AUC - ROC Curve

- FPR - False Positive Rate
  - The large errors are being penalised, yet the scale of the error is closer to that of mean absolute error..
- TPR - Total Positive Rate
  - The large errors are being penalised, yet the scale of the error is closer to that of mean absolute error.

- Steps to obtain TPR and FPR:

    1. Arrange the predicted probabilities in descending order.
    2. Take the first value as a threshold and the classes are predicted based on this threshold.

- Steps to obtain AUC - ROC :

    1. Set each of these predicted probability values as a threshold and generate the TPR and FPR
    2. Take the first value as a threshold and the classes are predicted based on this threshold.
    3. calculate TPR and FPR

4 Result and Conclusion

In this paper we Have designed a churn prediction model based on logistic regression Hence, this research aimed to build a system that predicts the churn of customers in Banks. These prediction models need to achieve high AUC values. To test and train the model, the sample data is divided into 80 percent for training and 2 percent for testing.We got the AUC- ROC as 0.7355364622101287 that reflects that our predictions gave a better result.

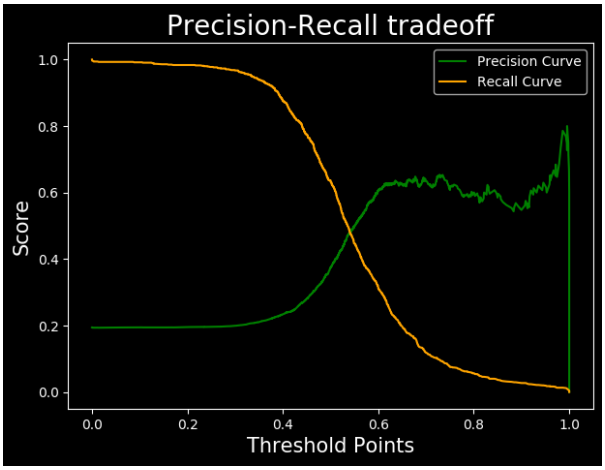

Figure 16. Prediction Tradeoff curve

5 Acknowledgement

We would like to Thank prof. Ritesh Kumar Mishra (Prof. NIT Patna) who had been constantly supporting and mentoring us and also helped shape our report which helped in giving us much deeper insights into the course which eventually helped us in the completion of this report.

	precision	recall	f1-score	support
0	0.89	0.75	0.81	3559
1	0.38	0.64	0.47	855
accuracy			0.72	4414
macro avg	0.64	0.69	0.64	4414
weighted avg	0.79	0.72	0.75	4414

Figure 17. Precision Recall Table

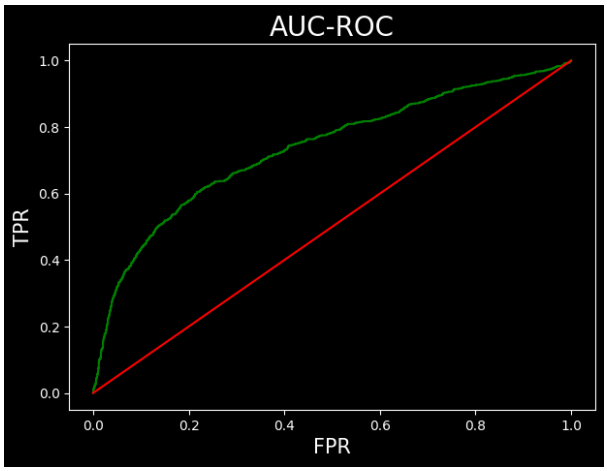

Figure 18. Resultant AUC-ROC Curve

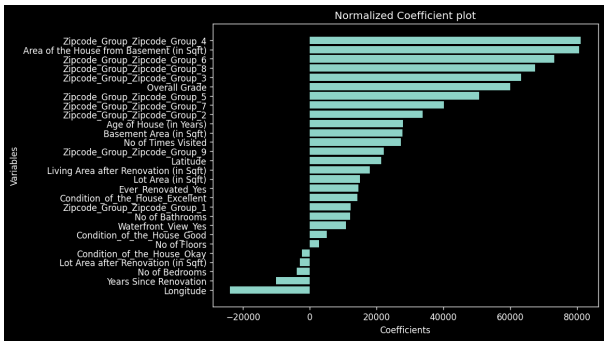

Figure 19. Model Coefficients

References

[1] https*https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5*

[2] https://miro.medium.com/max/1192/1*wilGXrItaMAJmZNl6RJq9Q

[3] https://arxiv.org/ftp/arxiv/papers/1904/1904.00690.pdf/

[4] https://www.geeksforgeeks.org/understanding-logistic-regression/

[5] https://thesai.org/Downloads/Volume11No5/Paper$_6$7−$Customer_Churn_Prediction_Model.pdf$

[6] https://trainings.internshala.com/machine-learning-training/?tracking$_source = trainings - dropdown - data - science - hp$