Customer churn prediction

Ta Duy Hai

February 23, 2024

Content

Content

Business Problem

Customer churn prediction is a critical aspect of business management. It involves understanding and addressing customer attrition, which refers to the loss of clients or customers.

For businesses in these sectors, measuring customer attrition is a vital business metric. This is because retaining an existing customer is significantly more cost-effective than acquiring a new one.

Content

Technical Problem

The task of predicting customer churn can be framed as a binary classification problem with two labels: 0 and 1. Here, label 0 represents customers who have not churned, while label 1 denotes churned customers.

Apart from demanding reasonable accuracy, classification models for customer churn prediction need to be interpretable to understand customer behavior and devise effective retention strategies.

Content

Model Evaluation Criteria

- Accuracy: Measures the overall correctness of the model's predictions.
- Precision: Indicates the proportion of correctly predicted positive cases out of all predicted positive cases.
- Recall: Measures the proportion of correctly predicted positive cases out of all actual positive cases.

Content

Data

A dataset relating characteristics of telephony account features and usage and whether or not the customer churned:

- churn-bigml-80: This dataset comprises characteristics of telephony account features and usage, coupled with information about customer churn. It serves as a fundamental component for training churn prediction models.
- churn-bigml-20 Dataset: Similar to the churn-bigml-80 dataset, this subset contains telephony account attributes and usage details, along with churn labels. It is utilized for testing and evaluating the performance of churn prediction models.

Data

This dataset consists of 20 fields of information:

State	Total evel call
Account length	Total eve charge
Area code	Total night minutes
International plan	Total night call
Voice mail plan	Total night charge
Number vmail message	Total intl minutes
Total day minutes	Total intl call
Total day calls	Total intl charge
Total eve minutes	Customer service calls
Total day charge	Churn

Data

The dataset contains text-type fields, so we need to use a label encoder to convert them into numerical values.

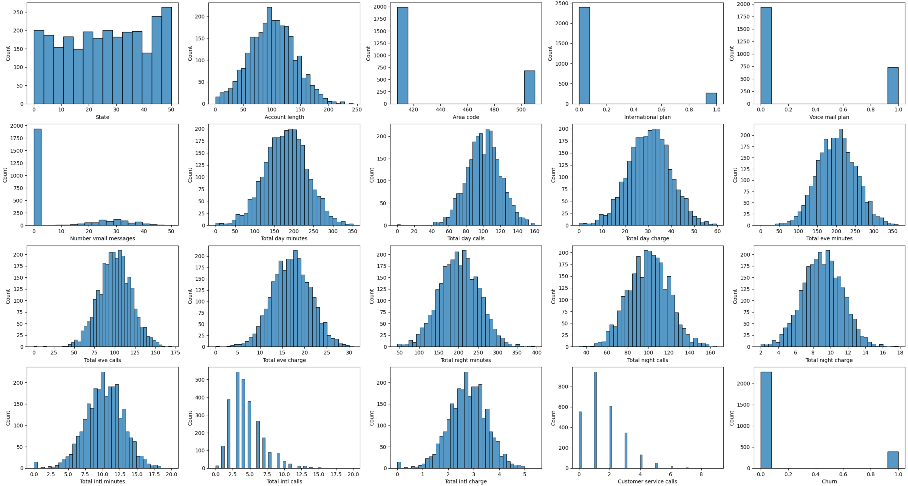

Figure: Histogram of train data.

Data

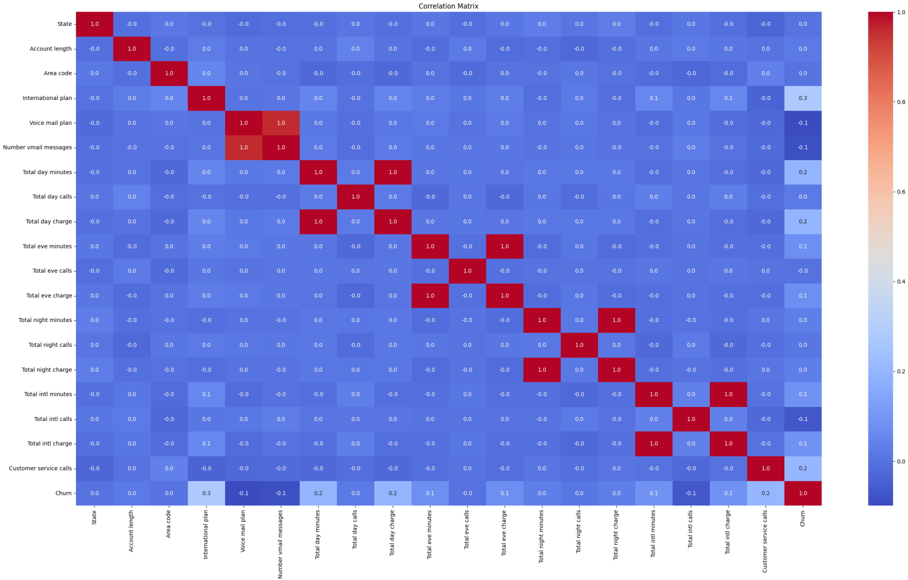

Figure: Correlation coefficient matrix of feature.

Data

From the initial training dataset, we will create two additional datasets:

- A dataset with highly correlated variables removed.
- A dataset with highly correlated variables removed and balanced labels for the churn attribute.

Content

Logistic Regression

Logistic Regression Model is a statistical model used to predict the probability of a binary dependent variable (e.g., Yes/No, Churn/Retained) based on a set of independent variables.

The model often uses the Sigmoid function to describe the relationship between the independent variables and the dependent variable.

The model is typically trained by minimizing the cross-entropy loss function.

Logistic Regression

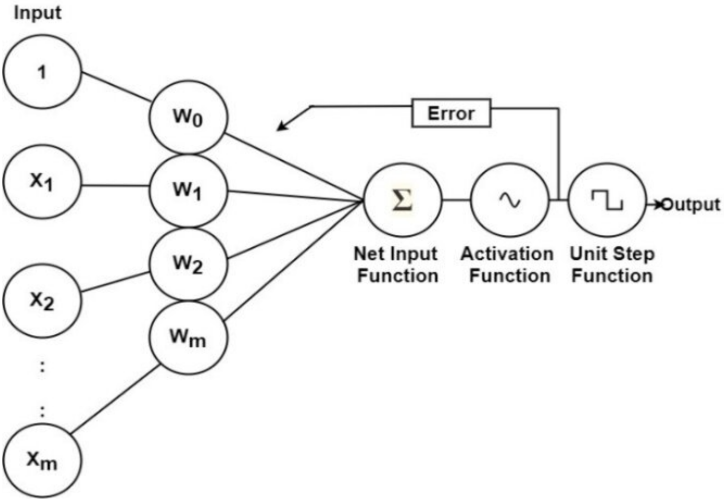

Figure: Architecture of Logistic Regression.

Decision Tree

Decision tree is a machine learning model that can be used for both classification and regression tasks.

It works by recursively splitting the data into smaller and smaller subsets until each subset contains only one type of data point (in the case of classification) or a specific value (in the case of regression.

A decision tree consists of nodes and edges, each node represents a feature, each edge represents a rule decision, each leave of the tree represent the final predictions.

Decision Tree

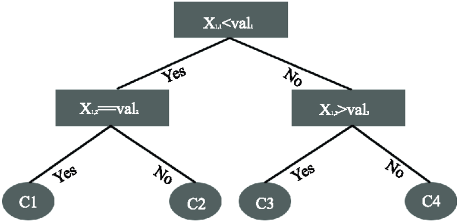

Figure: Architecture of Decision Tree.

Content

Logistic Regression

Data	Precision	Recall	Accuracy
Original	0.6611	0.3038	0.8890
Remove correlated variables	0.5044	0.3586	0.9017
Balanced and remove correlated variables	0.2709	0.7717	0.6481

Table: Result of Logistic Regression Models.

Decision Tree

Data	Precision	Recall	Accuracy
Remove correlated variables	0.8444	0.8	0.9505
Balanced and remove correlated variables	0.6451	0.8421	0.9115

Table: Result of Decision Tree Models.

Thank you for listening.