Predictive Modeling for Customer Churn Assignment Report

Objective:

The objective of this assignment is to build a predictive model that can predict customer churn for a given company. The intern will use machine learning techniques to build the model and document the process, including feature selection, model evaluation, and performance metrics.

Steps involved

Data preprocessing:

Dataset contains 4521 rows and 17 features.

Note: In the dataset it has mentioned that variable duration is highly correlated with output variable. Hence this feature has been removed to build a realistic predictive model

Data preprocessing involves following steps :

1.Data cleaning

- There are no missing values present in dataset.
- There are no duplicate values present in dataset.

2.Handling imbalanced data

- The output variable y has two outcomes(either ‘yes’ or ‘no’).In the given dataset 88% of output belongs to class ‘no’. Therefore the data is highly imbalanced dataset.
- I have used oversampling method to convert imbalanced dataset into balanced dataset.

3. Encoding the data

- By analysing the data I have found out that out of 16 features
    - 10 features were categorical features
    - 6 features were numerical features
- Now we have to convert categorical features into numerical features
    - Ordinal variable is converted into numerical feature based on mutual ordering among themselves.
    - For nominal features I have used one hot encoding to convert into numerical features.

After converting categorical variables to numerical variables are increased to 45 variables

4. Standardizing the dataset

This step is performed to :

- For faster convergence
- To ensure that variables measured at different scales do not contribute differently for analysis.

5.Feature selection

I have used mutual information from the field of information theory for feature selection. I have selected 15 best features among all the features using mutual information scores to use those features in prediction models.

Building prediction models

I have split data into 75% for training and 25% for testing ,after that I have applied various classification models on training and testing data. I have used three prediction models to predict the customer churn those models are:

- Logistic regression
- Gradient boosting
- Random forest classifier

Here is the table of results of all models which includes Accuracy ,precision, recall, F1 score and AUC score

Classifier	Accuracy	precision	Recall	F1 score	AUC score
Logistic regression	68%	62%	69%	65%	73%
Logistic regression (hyperparameter tunning)	68%	62%	69%	65%	73%
Gradient boosting	75%	69%	78%	73%	84%
Random forest classifier	98%	99%	96%	97%	100%

- From above table we can see that random forest classifier has given highest accuracy, precision,recall,F1 Score and AUC score.
- Therefore we can use Random forest to predict the customer churn.

The limitations of using a Random Forest model to predict customer churn are:

1. Random Forest models are prone to overfitting, which can lead to inaccurate predictions.

2. Random Forest models are computationally expensive and require a lot of time to train and test.

3. Random Forest models are difficult to interpret, as the resulting decision trees are often very complex and hard to understand.

To overcome these limitations, future work could include:

Exploring alternative models such as Support Vector Machines, Neural Networks, and Ensemble models. Additionally, feature selection and regularization techniques could be used to reduce overfitting and improve accuracy. Finally, data visualization techniques could be used to gain a better understanding of the data and the resulting decision trees.

Important links :

Python codes (Google colaboratory) link :

https://colab.research.google.com/drive/1YDdFYpztYWdBuuWcPVUG5UoeR9xd8rWx?usp=sharing

GitHub assignment link :

https://github.com/Ganeshwalimbe/Predictive-Modeling-for-Customer-Churn