REPORT

CUSTOMER LIFETIME VALUE PREDICTION

Introduction

Customer Lifetime Value represents a customer's value to a company over a period of time. It's a competitive market for insurance companies in 2019, and insurance premium isn't the only determining factor in a customer's decisions. CLV is a customer-centric metric, and a powerful base to build upon to retain valuable customers, increase revenue from less valuable customers, and improve the customer experience overall.

Project Overview

The objective of the problem is to accurately predict the Customer Lifetime Value(CLV) of the customer for an Auto Insurance Company.CLV is the total revenue the client will derive from their entire relationship with customer.

Problem statement

An Auto Insurance company in the USA is facing issues in retaining its customers and wants to advertise promotional offers for its loyal customers. They are considering Customer Lifetime Value CLV as a parameter for this purpose. Customer Lifetime Value represents a customer's value to a company over a period of time. It's a competitive market for insurance companies, and the insurance premium isn't the only determining factor in a customer's decisions. CLV is a customer-centric metric, and a powerful base to build upon to retain valuable customers, increase revenue from less valuable customers, and improve the customer experience overall. Using CLV effectively can improve customer acquisition and customer retention, prevent churn, help the company to plan its marketing budget, measure the performance of their ads in more detail, and much more.

Data collection and preprocessing

The dataset represents Customer lifetime value of an Auto Insurance Company in the United States, it includes over 24 features, it includes over 24 features and 9134 records to analyze the lifetime value of Customer.The data has a total of 9134 observation of 24 variables namely:customer state,customer lifetime value,response,coverage,education,effective to data,employment status,gender,income,location code,marital status,monthly premium auto,months since last claim,months since policy inception,number of open complaints,number of policies,policy type,renew offer type, sales channel,total claim amount,vehicle class,vehicle size.

Before training the machine learning models, we performed several preprocessing steps:

- Data Cleaning: We checked for missing values and removed any instances with incomplete information. - Feature Selection: We analyzed the relevance of each feature and selected the most informative ones for classification.

- Feature Scaling: To ensure all features were on a similar scale, we applied normalization or standardization techniques.

Exploratory Data Analysis

Performing EDA to understand the relation of target variable CLV with other features.

We are going to understand numerical features and the we also going to understand categorical features and the numerical attributes in proportion to each of the entries in a certain categorical feature or column.Categorical data refers to a data type that can be stored and identified based on the names or labels given to them.Numerical data refers to the data that is in the form of numbers,and not in any language or descriptive form.Also known as qualitative data as it qualifies data before classifying it.

Box plots,distribution plots and scatter plots are used in above features.

Univariate and Bivariate analysis takes place .In univariate Analysis we are going to understand the pattern of our variables and it looks at one variable,Bivariate Analysis looks at two variables and their relationship

Methodology

To predict and calculate CLTV,we have to estimate the frequency,recency and total amount of purchases by each customer.we are going to determine basic information about each customers average and lifetime purchase amount, as well as each customers duration and frequency of purchase.

Statistical Analysis techniques like OLS for numerical and Mann–Whitney U and also Kruskal Wallis test for the categorical variables were performed to find the significance of the features with respect to the target.

Supervised Regression Models like Linear Regression, Ridge Regression, Lasso Regression, DecisionTree Regression, Random Forest Regression and Adaboost Regression.

We use regression models because we had the problem statement about the prediction of the customer value.

We use regression models because we had the problem statement about the prediction of the customer value.

Using GridSearchCV with Random Forest Regression gave the best RMSE and R^2 score values

The decision of using an Anova or kruskal wallis test is the distribution of data.Normal/Gaussian distribution should be analysed with Anova while non-normal/non-gaussian distribution should be analysed with the kruskal wallis.

The major difference between the Mann-whitney U and the kruskal wallis H is simply that the latter can be accommodate more than two groups

The Anova(and t-test) is explicitly a test of equality if means of values.The kruskal-wallis(and Mannwhitney) can be seen technically as a comparison of the mean ranks.

Model Building and Evaluation

OLS model

OLS model is built for statistical anaylsis.

Training and testing data

Training data is subset of original data that is used to train the machine learning model,whereas testing data is used to check the accuracy of the model.

Train or test is a method to measure the accuracy of our model.It is called train or test because we split the data set into two sets:a training set and testing test.80% for training,and 20% for testing.we train the model using the training test and we test the model using testing set.

Checking Asssumptions about the data

-Normality :the errors will be normally distributed,by testing the residuals for a normality test the p value is more than 0.05.

-Multicollinearity:There should be no perfect linear relationship between two or more of the predictors.There is no multicollinearity in our data.

-Linearity:The relationship between X and the mean of Y is linear.

-Heteroscedasticity :The variance of the errors is not constant across observations.In particular the variance of the errors may be a function of explanatory variables.Hence, there is no Heteroscedacity in our data.

- No Auto correlation:It says that the error terms of different observation should not be correlated with each other.

Linear regression

Linear regression model is used.

Linear regression is a linear approach for modelling the relationship between input variables(x) and the single output variable(y).More specifically, that y can be calculated from a linear combination of the input variables.

Ridge and Lasso Regression

Ridge and lasso regression models were used.

Thus,ridge and lasso regression should be used when you are interested in optimizing for predictive ability rather than inference.

Decision Tree

Decision tree model is implemented.

Decision tree builds regression models in the form of a tree structure.It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed.The final result is a tree with decision nodes and leaf nodes.The output of a decision tree can also be easily understood.

Random Forest

Random forest regression algorithm is used.

Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression.Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make more accurate prediction than a single model.

Hyperparameter tuning of random forest

In Random forest, the hyperpparameters are the number of tress and the number of features and the type of trees(such as GBM OR M5). The number of features is important and should be tuned.in this case,random forest is useful because it automatically tunes the number of features.

RESULT

By OLS Regression mode:The R -squared value is 0.259, the adjusted R-squared value is 0.25

By Ridge and Lasso Regression model:

RMSE value is :0.5925342452447083, R-squared value:0.2148084953609085

RMSE value is :0.5992937645386806, R-squared value:0.19679167777854611

By decision tree :

RMSE:0.263,MAE:0.103,R-squared :0.844

By Random forest model:

RMSE:0.196,MAE:0.088,R-squared:0.913

By Hyperparamater tuning of random forest:

RMSE:0.197,MAE:0.089,R-squared:0.912.

Random forest model with hyper parameters tuning Using GridSearchCV gave the best RMSE and R^2 score values.

Conclusion

In all the model is a very good model apart from few of the assumptions not going in favour of the linear regression model guidelines namely heteroscedasticity.

*Overall we can see that a positive response, premium coverage, education of either high school or lower or masters, employed and retired, male gender, income, married or single status, monthly premium auto, no. of open complaints and no. of policies affect the CLV and has major role behind the values it comes up to

*Ironically being an auto insurance company the type of vehicle or size does not matter much

*Male gender has greater impact than the female indicating that male drivers are more

*Mainly the employed and the retired personnels need the insurance more and thus will contribute more to the company

*Singles/bachelors and married are the target customers as the former is more inclined towards fast driving and the latter a more towards work and responsibility commitment.

Future work

From the model we designed, we can suggest that

• The agents should target mainly the customers who are employed or retired and married or single and education is either very basic or master level

• The number of complaints should be reduced

• More attention should be given to the premium customers rather than the basic customers

• The target audience should be male

• The agents should start increasing their policy advertisement to the customers as the no. of policy affects the CLV

References

*The dataset represents Customer lifetime value of an Auto Insurance Company in the United States, it includes over 24 features, it includes over 24 features and 9134 records to analyze the lifetime value of Customer

*Hwang, Y. H. (2019). Hands-on data science for marketing: Improve your marketing strategies with machine learning using Python and R. Birmingham, UK: Packt Publishing.

*Jeffery, M. (2010). Data-driven marketing the 15 metrics everyone in marketing should know. Hoboken (N.J.), Canada: John Wiley.

* Müller, A. C., & Guido, S. (2018). Introduction to machine learning with Python: A guide for data scientists. Sebastopol, CA: O'Reilly Media.

*By browsing internet and referring you tube.