Predicting Customer Lifetime Value (CLV) of an Automobile Insurance Company using Various Customer Demographics

Jaswinder Singh

National College of Ireland
(M.Sc in Data Analytics)

x19219997@student.ncirl.ie

Video Presentation URL :
https://web.microsoftstream.com/video/c6b5743a-ca4b-4890-94ec-d1906bc97a58

Objectives of the Study

- **To predict the customer lifetime value (CLV) for individual customers of an automobile insurance company on the basis of different customer demographics.**
- **To conduct a qualitative and quantitative analysis of the results obtained.**

What does Customer Lifetime Value (CLV) mean?

- Customer lifetime value or CLV is a metric which is used to measure a customer's total worth to a business over the entire period of the customer's relationship with the company.
- CLV represents the total amount of money a customer is expected to spend in your business, or on your products, during their lifetime. It is important to know because it helps in making decisions about how much of the company's capital should be invested in acquiring new customers and also retaining existing ones.

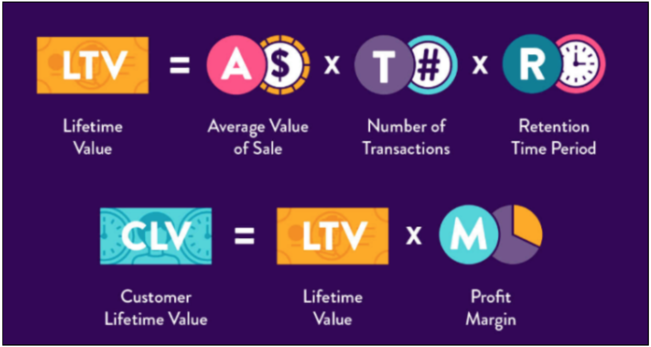
How to calculate customer lifetime value?

- There are many different ways in which one can calculate CLV for different consumers. In general, the CLV can be calculated by subtracting the total costs of the customer service, sales from the total revenue generated by the customer.

$$\mathrm{CLV} = \overbrace{\boxed{\sum_{t=1}^{T} \frac{\text{Revenue}}{(1+d)^t}}}^{\text{Revenue generated by the customer}} - \overbrace{\boxed{\sum_{i=1}^{T} \frac{\text{cost}}{(1+d)^t}}}^{\text{Cost of the customer services}} \tag{1}$$

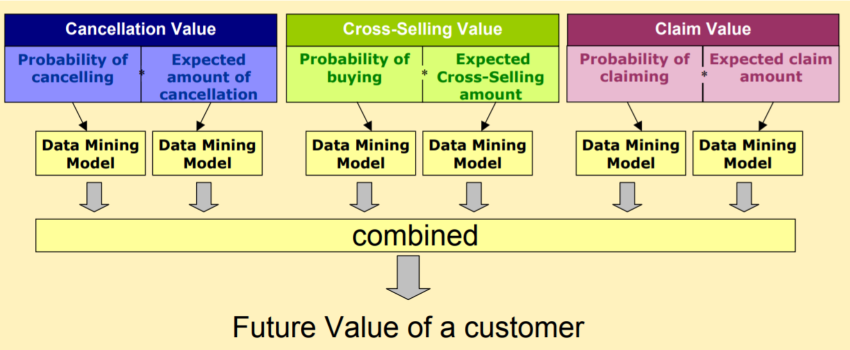
How to calculate customer lifetime value?

- Another way to calculate customer lifetime value is to multiply the lifetime value with the profit margin of the consumer.



Methodology Used for CLV Prediction

- We will be using the data mining based CRISP DM methodology for building a predictive model for CLV prediction. For predicting CLV, we have to take into account the following 3 components:
  - Value of claim
  - Value of cancellation
  - Value of cross-selling



Data Description

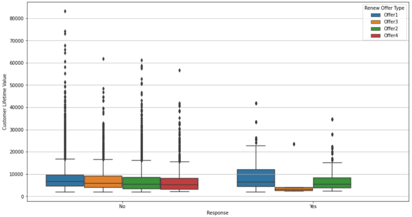
The data is acquired from kaggle. The data contains both categorical and numerical type variables.

Column_Name	Type	Description
Customer Lifetime Value	String	Customer's total worth to business over lifetime of the relationship
State	String	State of residence or business of the customer
Customer	Float	Customer ID No.
Response	String	Yes or No response to a renewal offer
Coverage	String	Type of Policy (Basic, Extended, Premium)
Education	String	Level of education of customer (High School or less, College, BA, MA, PhD)
Effective To Date	Date-time	Date on which the policy expires
EmploymentStatus	String	Employment Status of Customer (Employed, Unemployed, Retired, Disabled, Medical Leave)
Gender	String	Gender of customer (Male or Female)
Income	Integer	Customer's annual income
Location Code	String	Location type of customer (Urban, Rural, Suburban)
Marital Status	String	Marital Status of the customer (Married, Single, Divorced)
Monthly Premium Auto	Integer	Amount of customer's monthly insurance payments
Months Since Last Claim	Integer	Number of months between customer's last reported insurance claim
Months Since Policy Inception	Integer	Number of months since customer began an insurance policy
Number of Open Complaints	Integer	Number of unresolved customer complaints
Number of Policies	Integer	Number of policies customer currently owns
Policy Type	String	Type of policy (Corporate auto, Personal auto, Special auto)
Policy	String	3 levels (L1, L2, L3) as per policy type (Corporate, Personal, Special)
Renew Offer Type	String	4 types of renewal offers (Offer 1, Offer 2, Offer 3, Offer 4)
Sales Channel	String	Channels to purchase a policy (Agent, Branch, Call center, Web)
Total Claim Amount	Float	Cummulative amount of claims since policy inception
Vehicle Class	String	Type of vehicle (4-Door, Luxury, Luxury SUV, Sports car, SUV, 2-Door)
Vehicle Size	String	Size of vehicle (Large, Midsize, Small)

- The target variable in our data is skewed with a skewness of 3.03 which indicates presence of outliers in the data. Therefore, we apply log transformation to our target variable.

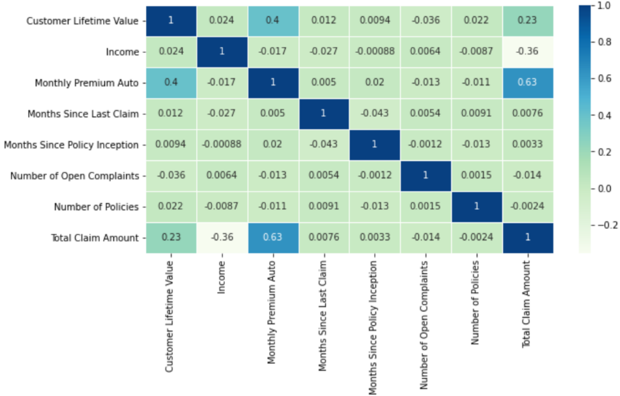

```
skewness before transformation: 3.0322802599299847
skewness after transformation: 0.5761717967456804
```

- Removal of Outliers:



Data Pre-Processing and Transformation

- As we can see from the correlation heat map, there are no two variables who have a high correlation.

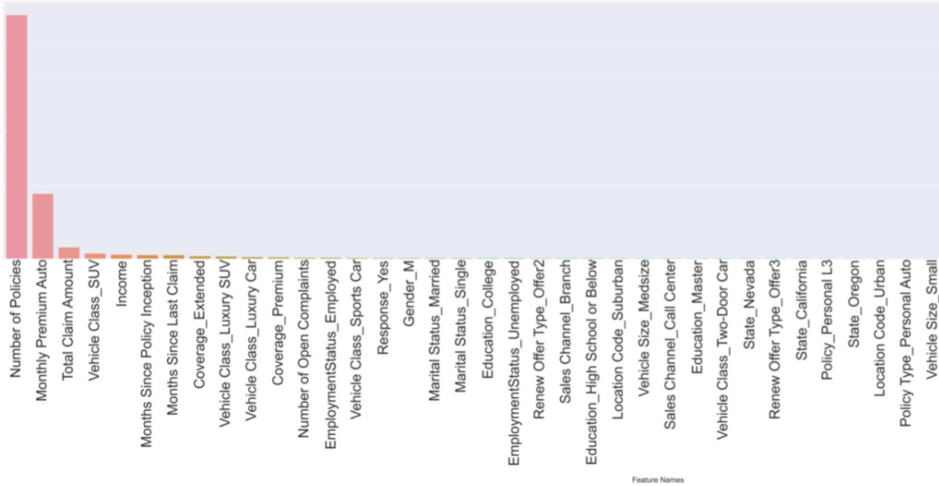


Model Implementation

- A random forest regressor model was chosen for implementation after critical review of the existing literature and understanding the complexity and size of the chosen data. The hyperparameter optimization was performed on the important features.
- The evaluation was done on the basis of R^2, adjusted R^2 and mean absolute percentage error (MAPE) values.

```
model: Random Forest
train r2: 0.9873330649917157
test r2: 0.9101948343426243
Mape: 0.9811802204788764
Adjusted R squared score is : 0.9085255933452753
------------------------------------------------------------
```

Feature Importance



Conclusion

The random forest regressor was then implemented which was able to achieve and R^2 of 0.98 for the train data and 0.91 for the test data. The adjusted R^2 value comes to be 0.908 which can be considered as an optimum value based on the given data. The model was able to achieve a mean absolute percentage error of just 0.98%. Overall, we can infer that our model has performed quite well to accomplish the given predictive analysis task.

The End