Customer Lifetime Value Prediction

Problem Statement: To predict the Customer Lifetime Value for an insurance company offering vehicle insurance.

Customer Lifetime Value is a generally used metric by companies and fiscal institutions to assign a numeric value to their guests and thereby inform their strategy of adding the companies ’ gains.

Companies generally use the formula CLV is = (Annual revenue per customer x Customer relationship in years) – Customer acquisition cost

When it comes to insurance, clients can be divided into various groups. Businesses create distinct policies because not all client segments will desire the same one. While some consumers might choose for less coverage, others might choose for more. This does not imply that consumers with less coverage are worth less to the business, since we also need to take their acquisition costs into account.

In order to choose which consumer segment to target, the insurance firm must analyze its current clientele while taking all of these variables into account.

The dataset includes historical information about the clients the business has already acquired, and the CLV has been calculated for each of them.

The objective is to create a model that can forecast the target variable by establishing the relationship between the explanatory variables and the target variable.

Dataset Introduction

This is Dataset is derived from kaggle. It is a insurance company dataset. It consists of 9134 Rows and 24 variables. CLV being the target/ response variable. This is the list of all the variables.

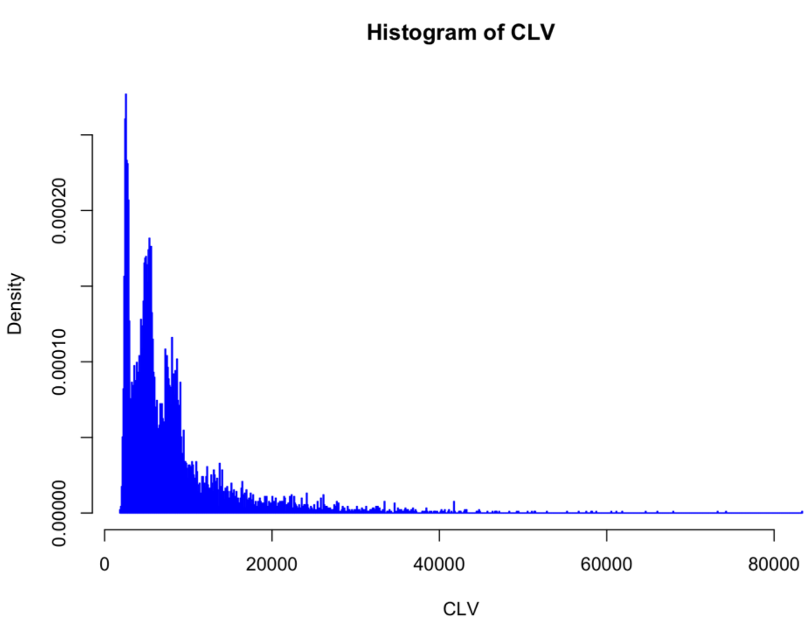
Customer	Unique ID assigned to customers
State	State to which customers belong
Customer Lifetime Value	Net profit generated by customers for the firm
Response	Positive or negative response with regards to purchase of policy plans
Coverage	Policy coverage chosen by the customers
Education	Education received by the customers
Effective To Date	Maturity date of insurance policy plan
EmploymentStatus	Customers' current employment status
Gender	Gender of customers
Income	Income level of cusomers
Location Code	Type of residential area of cusomers
Marital Status	Relationship status
Monthly Premium Auto	Monthly premium paid for the policy
Months Since Last Claim	Number of months that passed since the last claim made by the customer
Months Since Policy Inception	Number of months since the activation of policy plan
Number of Open Complaints	Number of unsolved complaints made by the customer
Number of Policies	Total number of policies purchased
Policy Type	Type of policy under the main categories
Policy	Category of policy plan adopted by the customer - personal, corporate or special
Renew Offer Type	Class of renewal offer accepted by the customer
Sales Channel	Channel via sales with a particular customer occurred
Total Claim Amount	Total amount that can be claimed by the customer on/before policy maturity
Vehicle Class	Class to which the insured vehicle belongs
Vehicle Size	Size of the customers' insured vehicle

The image above shows the variables of the dataset. The data consists of 16 categorical variables and 8 Numerical variables.

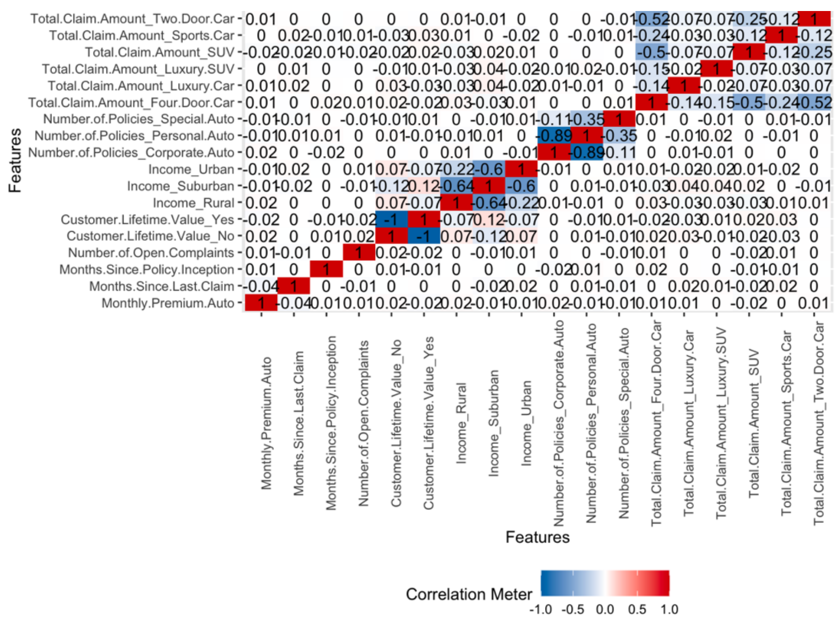
Data Summary

```
> df <- read_excel("Desktop/CLV.xlsx")
> summary(df)
   Customer            State             Response           Coverage          Education         Effective.To.Date 
 Length:9134        Length:9134        Length:9134        Length:9134        Length:9134        Length:9134       
 Class :character   Class :character   Class :character   Class :character   Class :character   Class :character  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character  
                                                                                                                  
                                                                                                                  
                                                                                                                  
 EmploymentStatus      Gender              Income      Location.Code      Marital.Status     Monthly.Premium.Auto
 Length:9134        Length:9134        Min.   :    0   Length:9134        Length:9134        Min.   : 61.00      
 Class :character   Class :character   1st Qu.:    0   Class :character   Class :character   1st Qu.: 68.00      
 Mode  :character   Mode  :character   Median :33890   Mode  :character   Mode  :character   Median : 83.00      
                                       Mean   :37657                                         Mean   : 93.22      
                                       3rd Qu.:62320                                         3rd Qu.:109.00      
                                       Max.   :99981                                         Max.   :298.00      
 Months.Since.Last.Claim Months.Since.Policy.Inception Number.of.Open.Complaints Number.of.Policies
 Min.   : 0.0            Min.   : 0.00                 Min.   :0.0000            Min.   :1.000     
 1st Qu.: 6.0            1st Qu.:24.00                 1st Qu.:0.0000            1st Qu.:1.000     
 Median :14.0            Median :48.00                 Median :0.0000            Median :2.000     
 Mean   :15.1            Mean   :48.06                 Mean   :0.3844            Mean   :2.966     
 3rd Qu.:23.0            3rd Qu.:71.00                 3rd Qu.:0.0000            3rd Qu.:4.000     
 Max.   :35.0            Max.   :99.00                 Max.   :5.0000            Max.   :9.000     
 Policy.Type           Policy          Renew.Offer.Type   Sales.Channel      Total.Claim.Amount Vehicle.Class     
 Length:9134        Length:9134        Length:9134        Length:9134        Min.   :   0.099   Length:9134       
 Class :character   Class :character   Class :character   Class :character   1st Qu.: 272.258   Class :character  
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median : 383.945   Mode  :character  
                                                                             Mean   : 434.089                     
                                                                             3rd Qu.: 547.515                     
                                                                             Max.   :2893.240                     
 Vehicle.Size       Customer.Lifetime.Value
 Length:9134        Min.   : 1898          
 Class :character   1st Qu.: 3994          
 Mode  :character   Median : 5780          
                    Mean   : 8005          
                    3rd Qu.: 8962          
                    Max.   :83325          
```

The summary abv shows some of the averages, ranges, median, 1st and 3rd quartile for the variables. For instance, average income for the customers is $37657. And the average CLV as of the current data is $8005.



The distribution is highly skewed, as seen by this figure, which suggests that the vast majority of customers have a lower customer lifetime value for the business. Very few clients fall into the greater lifetime value category.
The company's "ideal" consumers are few in number, so in order for them to generate revenue, they must also concentrate on serving the larger number of customers who have lower CLV.



The correlation plot indicates the correlation between the features in the dataset. Key observations include:

- The relationship between Total Claim Amount_SUV and Total Claim Amount_Luxury.Car and Customer Lifetime Value is favorable.
- There is a negative correlation between Months and Monthly Premium Auto.Since the previous claim.
- Total Claim Amount_SUV and Total Claim Amount_Luxury.Car have a favorable correlation with Income Urban.

Linear Regression Model

```
Call:
lm(formula = Customer.Lifetime.Value ~ State + Response + Coverage +
    Education + EmploymentStatus + Gender + Income + Location.Code +
    Marital.Status + Months.Since.Last.Claim + Months.Since.Policy.Inception +
    Number.of.Open.Complaints + Number.of.Policies + Policy +
    Renew.Offer.Type + Sales.Channel + Total.Claim.Amount + Vehicle.Class +
    Vehicle.Size, data = train)

Residuals:
   Min     1Q Median     3Q    Max
-11351  -3277  -1387    798  64097
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                    5.916e+03  7.416e+02   7.978 1.75e-15 ***
StateCalifornia                7.560e+01  2.298e+02   0.329  0.74218
StateNevada                    1.158e+02  3.190e+02   0.363  0.71665
StateOregon                    1.712e+02  2.383e+02   0.718  0.47258
StateWashington                3.387e+02  3.266e+02   1.037  0.29978
ResponseYes                   -2.056e+02  2.514e+02  -0.818  0.41351
CoverageExtended               1.446e+03  1.862e+02   7.766 9.41e-15 ***
CoveragePremium                3.639e+03  3.170e+02  11.480  < 2e-16 ***
EducationCollege               6.622e+01  2.083e+02   0.318  0.75054
EducationDoctor               -6.711e+01  4.286e+02  -0.157  0.87557
EducationHigh School or Below  4.099e+02  2.108e+02   1.945  0.05186 .
EducationMaster                1.771e+02  3.207e+02   0.552  0.58081
EmploymentStatusEmployed       5.607e+02  4.272e+02   1.313  0.18939
EmploymentStatusMedical Leave  2.337e+02  5.319e+02   0.439  0.66046
EmploymentStatusRetired       -4.497e+02  6.016e+02  -0.748  0.45475
EmploymentStatusUnemployed    -2.989e+02  4.298e+02  -0.695  0.48677
GenderM                       -3.661e+02  1.612e+02  -2.270  0.02322 *
Income                         1.169e-04  4.673e-03   0.025  0.98004
Location.CodeSuburban         -2.995e+01  3.163e+02  -0.095  0.92456
Location.CodeUrban             1.701e+02  2.925e+02   0.582  0.56085
Marital.StatusMarried         -3.283e+02  2.349e+02  -1.398  0.16225
Marital.StatusSingle          -7.493e+02  2.715e+02  -2.760  0.00579 **
Months.Since.Last.Claim        8.437e+00  7.982e+00   1.057  0.29052
Months.Since.Policy.Inception  1.278e+00  2.895e+00   0.441  0.65895
Number.of.Open.Complaints     -2.846e+02  8.795e+01  -3.236  0.00122 **
Number.of.Policies             5.911e+01  3.362e+01   1.758  0.07872 .
PolicyCorporate L2            -8.457e+02  5.145e+02  -1.644  0.10031
PolicyCorporate L3            -3.505e+02  4.703e+02  -0.745  0.45610
PolicyPersonal L1             -4.342e+02  4.628e+02  -0.938  0.34814
PolicyPersonal L2             -2.007e+01  4.393e+02  -0.046  0.96355
PolicyPersonal L3             -1.764e+02  4.268e+02  -0.413  0.67942
PolicySpecial L1               2.149e+02  9.936e+02   0.216  0.82876
PolicySpecial L2               1.644e+02  7.373e+02   0.223  0.82362
PolicySpecial L3               8.355e+02  7.442e+02   1.123  0.26158
Renew.Offer.TypeOffer2        -9.574e+02  2.012e+02  -4.758 2.00e-06 ***
Renew.Offer.TypeOffer3        -6.231e+02  2.441e+02  -2.553  0.01071 *
Renew.Offer.TypeOffer4        -1.323e+03  2.810e+02  -4.710 2.53e-06 ***
Sales.ChannelBranch            1.691e+02  1.999e+02   0.846  0.39753
Sales.ChannelCall Center       2.387e+02  2.276e+02   1.049  0.29414
Sales.ChannelWeb              -7.044e+01  2.478e+02  -0.284  0.77620
Total.Claim.Amount             3.509e-01  5.673e-01   0.619  0.53623
Vehicle.ClassLuxury Car        1.015e+04  7.216e+02  14.068  < 2e-16 ***
Vehicle.ClassLuxury SUV        1.050e+04  7.192e+02  14.595  < 2e-16 ***
Vehicle.ClassSports Car        3.456e+03  3.858e+02   8.958  < 2e-16 ***
Vehicle.ClassSUV               3.711e+03  2.454e+02  15.123  < 2e-16 ***
Vehicle.ClassTwo-Door Car      1.237e+02  2.078e+02   0.595  0.55166
Vehicle.SizeMedsize            4.147e+02  2.667e+02   1.555  0.11997
Vehicle.SizeSmall              3.648e+02  3.112e+02   1.172  0.24125
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6370 on 6345 degrees of freedom
Multiple R-squared:  0.1687,    Adjusted R-squared:  0.1626
F-statistic: 27.41 on 47 and 6345 DF,  p-value: < 2.2e-16
```

The model displays important coefficients:

The baseline value, or $5,916, is represented by the intercept when all predictors are zero.

Nevada, Oregon, Washington, California, and Nevada all have different effects on the response variable.
While CoverageExtended and CoveragePremium have good impacts, ResponseYes has a negative influence.
There is a negative correlation with GenderM (Male).
Now for the important predictors:

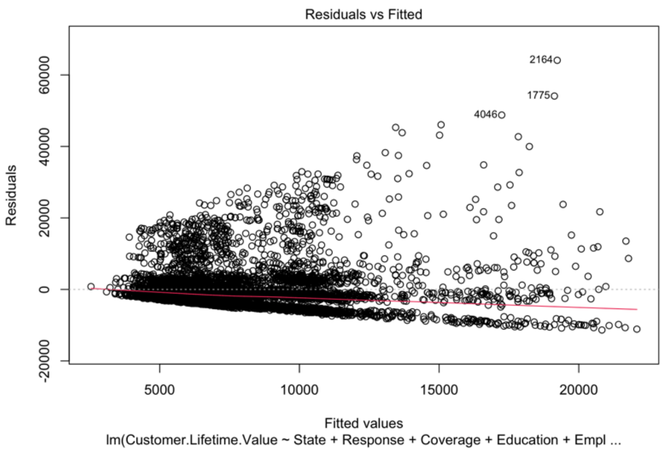
Vehicle Class, Renew Offer Type, and Number of Open Complaints are statistically significant.
These factors are essential in determining how we will react.

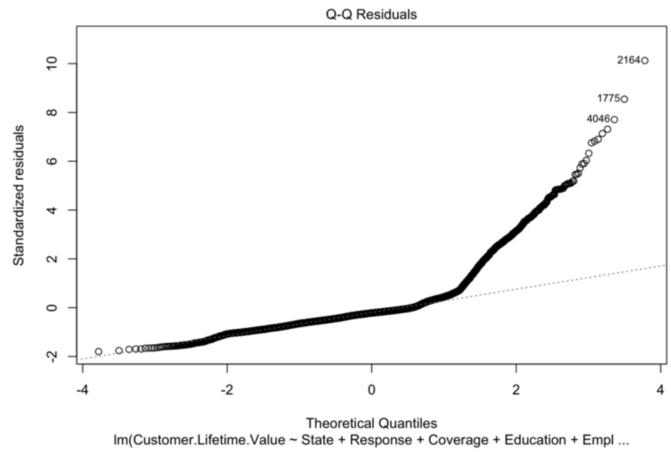
R-squared: The model explains 16.87% of the variation in the response variable.

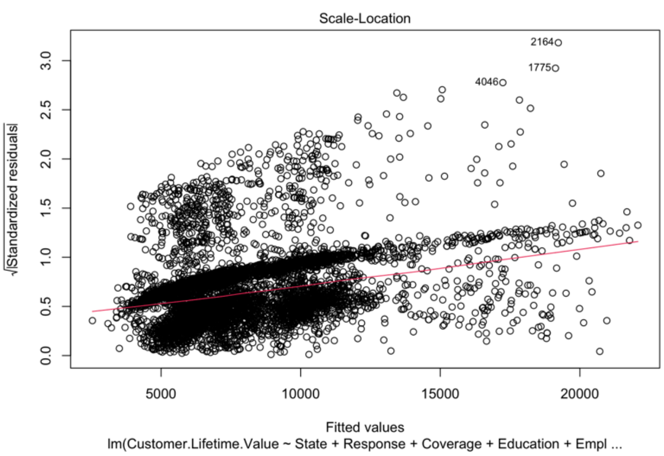
After controlling for predictors, the adjusted R-squared was 16.26%.
The remaining standard error The observed and projected values varied by an average of 6370.

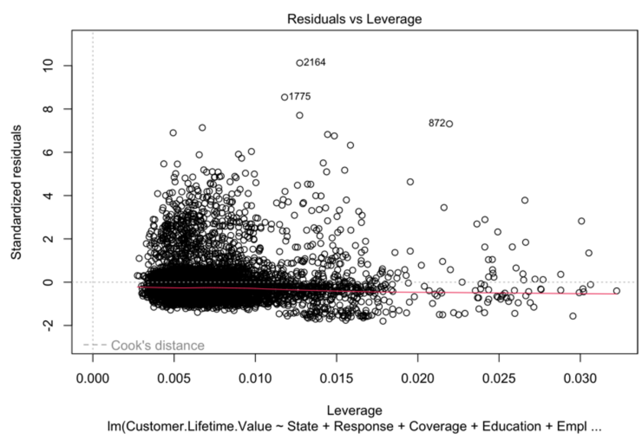
With a p-value of less than 2.2e-16 and an F-statistic of 27.41, our model's overall significance is demonstrated.

Residual Plots









4 plots were produced by the model:

1. There is no funnel form visible in the residues in the residual vs. fitted graph.

2. The Q-Q plot is followed by the points in the graphs' central region. There is a slight departure from the Q-Q plot in the trailing segment. But the leading part shows a large deviation from the Q-Q plot, suggesting that normalcy is not followed. The graph appears to be similar to the normal curve.

3. The residuals are distributed uniformly throughout the predictor ranges. A horizontal line with dots distributed equally (randomly) is seen.

4. The regression line is essentially straight, despite the appearance of extreme values.

Variable Selection: Backward Elimination

Backward elimination will be used to improve the model for further analysis in other models.

The final model is mentioned below.

- Customer.Lifetime.Value ~ Coverage + EmploymentStatus + Gender + Marital.Status + Number.of.Open.Complaints + Number.of.Policies + Renew.Offer.Type + Vehicle.Class
- Reducing the AIC from 112045.5 to 112014.3.

```
Step:  AIC=112014.3
Customer.Lifetime.Value ~ Coverage + EmploymentStatus + Gender +
    Marital.Status + Number.of.Open.Complaints + Number.of.Policies +
    Renew.Offer.Type + Vehicle.Class

                            Df  Sum of Sq        RSS    AIC
<none>                                    2.5848e+11 112014
- Number.of.Policies         1 1.2263e+08 2.5861e+11 112015
- Marital.Status             2 2.4577e+08 2.5873e+11 112016
- Gender                     1 1.9798e+08 2.5868e+11 112017
- Number.of.Open.Complaints  1 4.3436e+08 2.5892e+11 112023
- EmploymentStatus           4 8.4461e+08 2.5933e+11 112027
- Renew.Offer.Type           3 1.3350e+09 2.5982e+11 112041
- Coverage                   2 8.2783e+09 2.6676e+11 112212
- Vehicle.Class              5 3.6487e+10 2.9497e+11 112848
```

In Sample and Out of Sample Predictions

- In sample MSE: 40580557
  - 40580557 suggests the model's predictions within the training data have, on average, squared differences of approximately 40.6 million.

- Out of Sample MSE: 37722300
  - The value of 37722300 indicates that, on average, the model's predictions on testing data have squared differences of approximately 37.7 million.

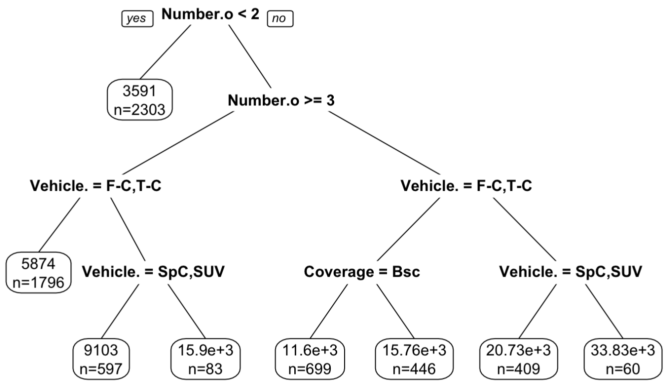
- The out-of-sample MSE is slightly lower than the in-sample MSE, suggests that the model is not overfitting to the training data and is performing reasonably well on new data.

Regression Tree Model

```
> # Model Building: Fiting regression tree
> insurance_rpart <- rpart(formula = Customer.Lifetime.Value ~ Coverage + EmploymentStatus + Marital.Status +
+                            Months.Since.Last.Claim + Number.of.Open.Complaints + Number.of.Policies +
+                            Renew.Offer.Type + Total.Claim.Amount + Vehicle.Class, data = train)
>
> # Printing and plotting the tree
> insurance_rpart
n= 6393

node), split, n, deviance, yval
      * denotes terminal node

 1) root 6393 309752900000  8011.663
   2) Number.of.Policies< 1.5 2303   4429787000  3590.690 *
   3) Number.of.Policies>=1.5 4090 234965500000 10501.030
     6) Number.of.Policies>=2.5 2476  15525540000  6989.099
      12) Vehicle.Class=Four-Door Car,Two-Door Car 1796   2380353000  5874.486 *
      13) Vehicle.Class=Luxury Car,Luxury SUV,Sports Car,SUV 680   5020707000  9932.986
        26) Vehicle.Class=Sports Car,SUV 597   1218461000  9102.869 *
        27) Vehicle.Class=Luxury Car,Luxury SUV 83    431830400 15903.830 *
     7) Number.of.Policies< 2.5 1614 142054100000 15888.600
      14) Vehicle.Class=Four-Door Car,Two-Door Car 1145  52757140000 13219.610
        28) Coverage=Basic 699  21436380000 11600.670 *
        29) Coverage=Extended,Premium 446  26617420000 15756.900 *
      15) Vehicle.Class=Luxury Car,Luxury SUV,Sports Car,SUV 469  61227740000 22404.580
        30) Vehicle.Class=Sports Car,SUV 409  39798810000 20727.810 *
        31) Vehicle.Class=Luxury Car,Luxury SUV 60  12440430000 33834.500 *
> prp(insurance_rpart,digits = 4, extra = 1)
```

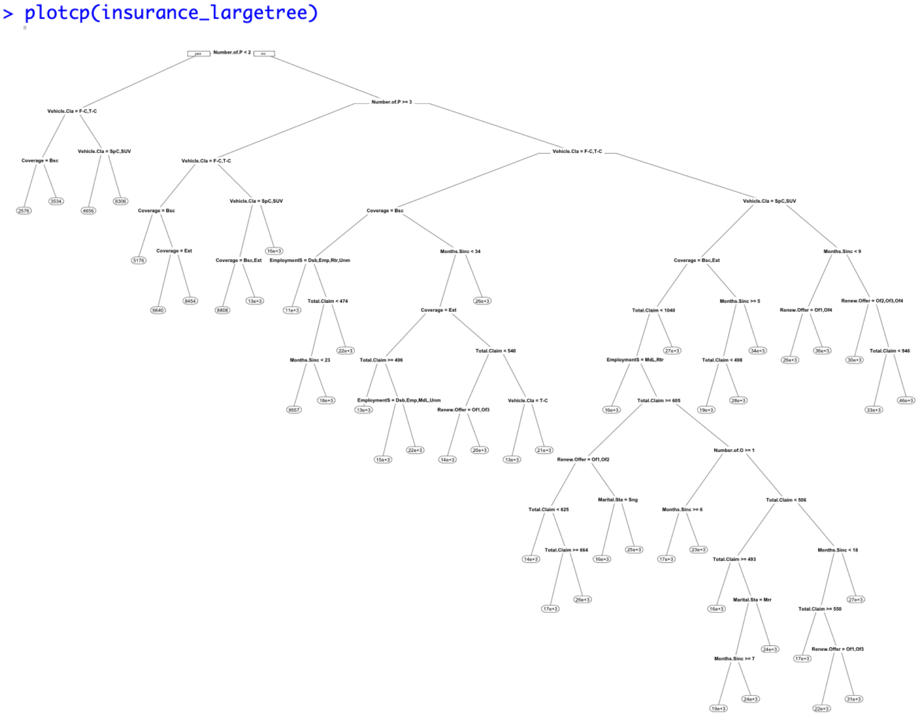


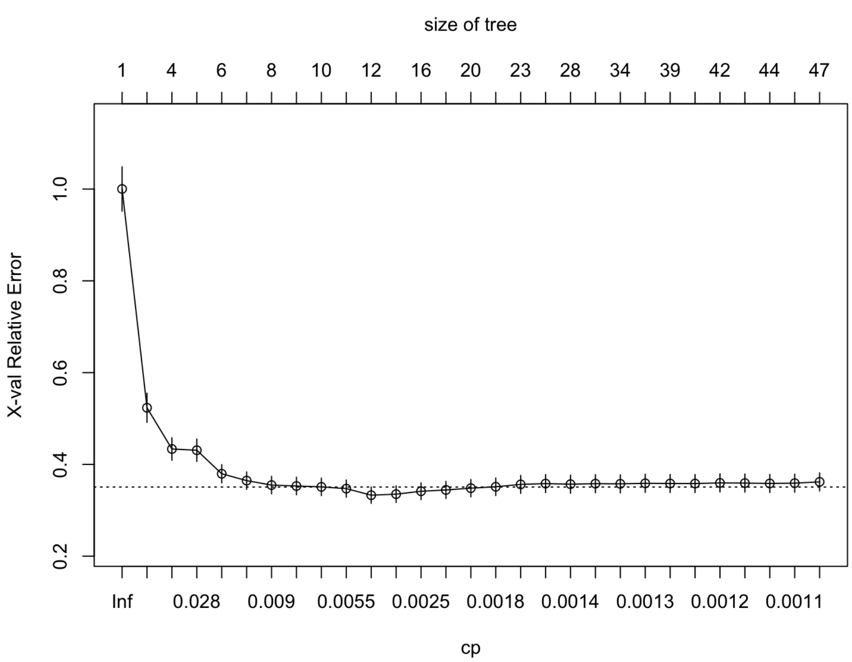
Interpretation:
The code is for building a decision tree model using the "rpart" package in R for predicting the "Customer.Lifetime.Value" based on various predictor variables.

1. **Root Node (Node 1):**
    - The initial node represents the entire dataset with 6393 observations.
    - The deviance is 309752900000, and the mean predicted value is 8011.663.
2. **Split on "Number.of.Policies":**
    - If the number of policies is less than 1.5, the model reaches Terminal Node 2.
    - If the number of policies is 1.5 or greater, it proceeds to Node 3.
3. **Node 3 - Further Split on "Number.of.Policies":**
    - If the number of policies is 2.5 or more, the tree leads to Node 6.
    - If the number of policies is less than 2.5, the model navigates to Node 7.
4. **Node 6 - Split on "Vehicle.Class":**
    - If the vehicle class is a Four-Door Car or a Two-Door Car, the prediction goes to Terminal Node 12.
    - If the vehicle class is a Luxury Car, Luxury SUV, Sports Car, or SUV, the model proceeds to Node 13.
5. **Node 13 - Further Split on "Vehicle.Class":**
    - If the vehicle class is a Sports Car or an SUV, the prediction goes to Terminal Node 26.
    - If the vehicle class is a Luxury Car or a Luxury SUV, the model reaches Terminal Node 27.
6. **Node 7 - Split on "Vehicle.Class":**
    - If the vehicle class is a Four-Door Car or a Two-Door Car, the prediction goes to Node 14.
    - If the vehicle class is a Luxury Car, Luxury SUV, Sports Car, or SUV, the model proceeds to Node 15.
7. **Node 14 - Split on "Coverage":**
    - If the coverage is Basic, the prediction goes to Terminal Node 28.
    - If the coverage is Extended or Premium, the model reaches Terminal Node 29.
8. **Node 15 - Split on "Vehicle.Class":**
    - If the vehicle class is a Sports Car or an SUV, the prediction goes to Terminal Node 30.
    - If the vehicle class is a Luxury Car or a Luxury SUV, the model reaches Terminal Node 31.

The tree structure can be visualized using the **prp** function

```
> # In sample prediction
> insuarnce_train_pred_tree = predict(insurance_rpart)
>
> # Out of sample prediction
> insurance_test_pred_tree = predict(insurance_rpart,test)
>
> # Out of sample MSE
> MSE.tree <- mean((insurance_test_pred_tree - test$Customer.Lifetime.Value)^2)
> MSE.tree
[1] 15972289
>
> # Pruning
> # Generate large tree
> insurance_largetree <- rpart(formula = Customer.Lifetime.Value ~ Coverage + EmploymentStatus + Marital.Status +
+                              Months.Since.Last.Claim + Number.of.Open.Complaints + Number.of.Policies +
+                              Renew.Offer.Type + Total.Claim.Amount + Vehicle.Class, data = train, cp = 0.001)
> prp(insurance_largetree)
>
> # Plotting the cp values
> plotcp(insurance_largetree)
```





Interpretation:
Using a decision tree model, both in-sample and out-of-sample predictions, calculates the mean squared error (MSE) for each dataset, and explores pruning for tree optimization.

The average squared difference between the predicted and actual values for the training dataset is shown by the In-Sample MSE of 16769791.

The average squared difference between the test dataset's expected and actual values is 16485081, which is the Out-of-Sample MSE.

Optimal tree size is the first one to cross the dash line in the plot - the leftmost number to cross the line (0.009)
To plot the optimal model, use the optimal number we found from the cp plot.
A good choice of cp for pruning is often the leftmost value for which the mean lies below the horizontal line.

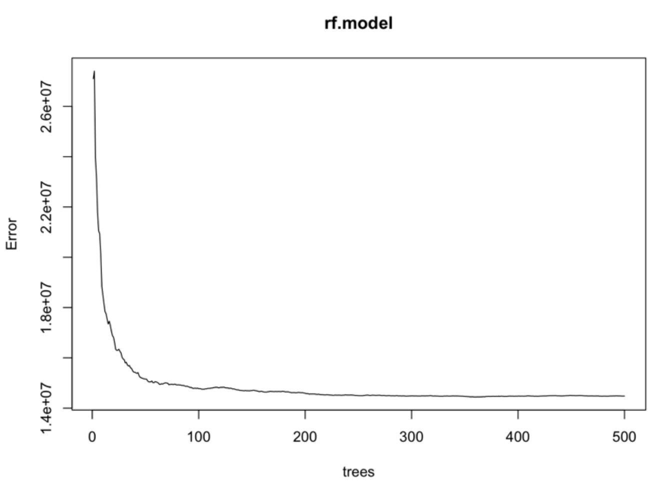
Random Forest Model

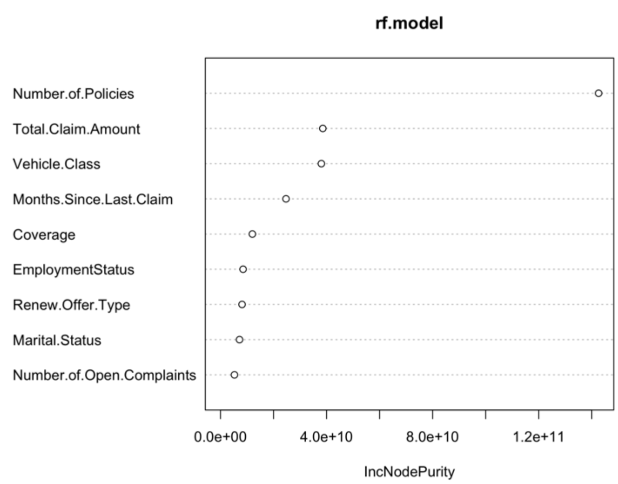
```
> # Random Forest
> rf.model <- randomForest(Customer.Lifetime.Value ~ Coverage + EmploymentStatus + Marital.Status +
+                          Months.Since.Last.Claim + Number.of.Open.Complaints + Number.of.Policies +
+                          Renew.Offer.Type + Total.Claim.Amount + Vehicle.Class, data = train, proximity = TRUE)
> rf.model

Call:
 randomForest(formula = Customer.Lifetime.Value ~ Coverage + EmploymentStatus +      Marital.Status + Months.Since.Last.Claim + Number.of.Open.Complaints +      Number.of.Policies + Renew.Offer.Type + Total.Claim.Amount +      Vehicle.Class, data = train, proximity = TRUE)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 3

          Mean of squared residuals: 14478759
                    % Var explained: 70.12
```

```
> # Find number of trees that produce lowest test MSE
> which.min(rf.model$mse)
[1] 360
>
> # Find RMSE of best model
> sqrt(rf.model$mse[which.min(rf.model$mse)])
[1] 3800.392
```





The model predicts Customer Lifetime Value based on predictors: Coverage, EmploymentStatus, Marital.Status, Months.Since.Last.Claim, Number.of.Open.Complaints, Number.of.Policies, Renew.Offer.Type, Total.Claim.Amount, Vehicle.Class.

Type of random forest: Regression
Indicates that this is a regression task, predicting a continuous numerical outcome (Customer Lifetime Value).
Number of trees: 500
The ensemble consists of 500 decision trees.
No. of variables tried at each split: 3
At each decision point in a tree, the algorithm considers 3 randomly selected predictors.

Model Performance:
Mean of squared residuals: 14478759
The average squared difference between predicted and actual values.

% Var explained: 70.12
The proportion of the response variable's volatility (Customer Lifetime Value) that the model can account for.
Shows the degree to which the model accurately represents the data's variability.

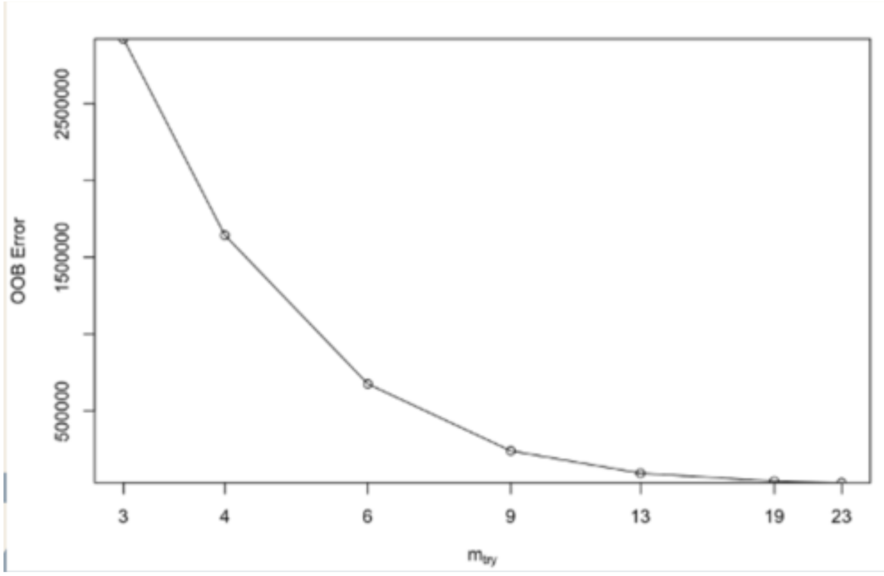
Model Evaluation:
Number of trees for lowest test MSE: 360
Indicates that the model achieves its lowest Mean Squared Error (MSE) on the test set with 360 trees.

Root Mean Squared Error (RMSE) of the best model: 3800.392
The square root of the MSE for the model with the lowest test MSE.
Represents the average absolute prediction error of the model.

Interpretation:
With a high proportion of variation explained (70.12%) and a reasonably low mean squared residual, the Random Forest regression model looks to be operating effectively.
Using 500 trees and experimenting with three variables at every split points to a strong ensemble model.
The best model, which has 360 trees, provides an estimate of the typical prediction error on the test set with a Root Mean Squared Error of 3800.392.
Based on the above predictor factors, the Random Forest model appears to be useful in estimating Customer Lifetime Value overall.

Model Tuning

```
> # Tune the model
> model_tuned <- tuneRF(
+   x=df[,-1], #define predictor variables
+   y=df$Customer.Lifetime.Value, #define response variable
+   ntreeTry=500,
+   mtryStart=4,
+   stepFactor=1.5,
+   improve=0.01,
+   trace=FALSE #don't show real-time progress
+ )
-0.7774307 0.01
0.5884241 0.01
0.6435207 0.01
0.604587 0.01
0.5434091 0.01
0.2281728 0.01
> plot(model_tuned)
```



Input Parameters:
Predictors (x): All columns in the dataframe df except the first one.
Response variable (y): Customer.Lifetime.Value from the dataframe df.
Number of Trees to Try (ntreeTry): 500.
Minimum Number of Variables for Split (mtryStart): 4.
Step Factor: 1.5.
Improvement Threshold (improve): 0.01.
Trace: Set to FALSE, meaning no real-time progress display.

It appears that the result is a set of numbers that show how well the tuned Random Forest model performed at various points during the tuning procedure.

Each Line:
An indicator of the model's performance (maybe out-of-bag error) appears in the first column.

The improvement threshold determined throughout the tuning procedure is represented by the value "0.01" in the second column.

Interpretation:
Through parameter adjustments, the Random Forest model is fine-tuned.
An indicator of the model's evolution can be seen in the series of performance numbers and improvements at each phase.

The numbers in the last column may help choose the best possible set of Random Forest model parameters. Depending on the parameters and improvement threshold provided, the optimal-tuned Random Forest model may be present in the model_tuned object.

The plot, generated by this function, shows the out-of-bag estimated error on the y-axis and the number of predictors utilized at each split when creating the trees on the x-axis.

Final Conclusion:

1. **Linear Regression Model:**
    - **R-squared:** 16.87%
    - **Residual Standard Error:** 6370
    - **In-Sample MSE:** 40580557
    - **Out-of-Sample MSE:** 37722300
2. **Regression Tree Model (Decision Tree):**
    - **In-Sample MSE:** 16769791
    - **Out-of-Sample MSE:** 16485081
    - (Additional information from a previous response: Root Node Mean: 8011.663)
3. **Random Forest Model:**
    - **Mean Squared Residuals:** 14,478,759
    - **% Variance Explained:** 70.12
    - **Optimal Trees for Lowest Test MSE:** 360
    - **RMSE of Best Model:** 3800.392

Conclusion:

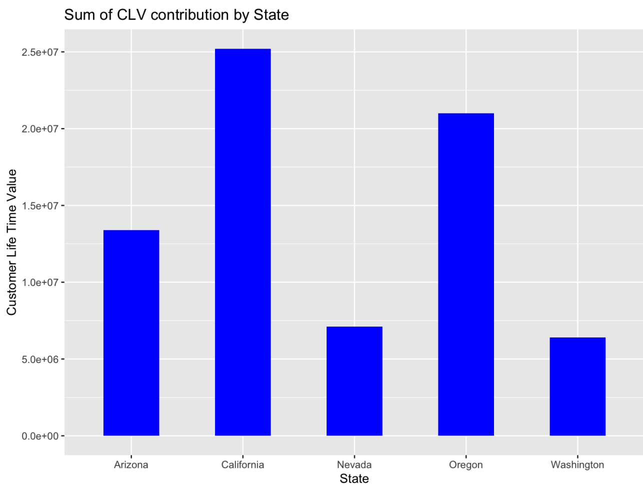
- The Linear Regression Model displays both in- and out-of-sample MSE values and performs moderately, with a rather low R-squared value.
- Both in-sample and out-of-sample predictions perform well with lower MSE values when using the Regression Tree Model (Decision Tree).
- The Random Forest Model has the highest percentage of Variance Explained among the models and determines the ideal number of trees for the lowest test MSE. The tree structure shows the hierarchy of requirements for making predictions.

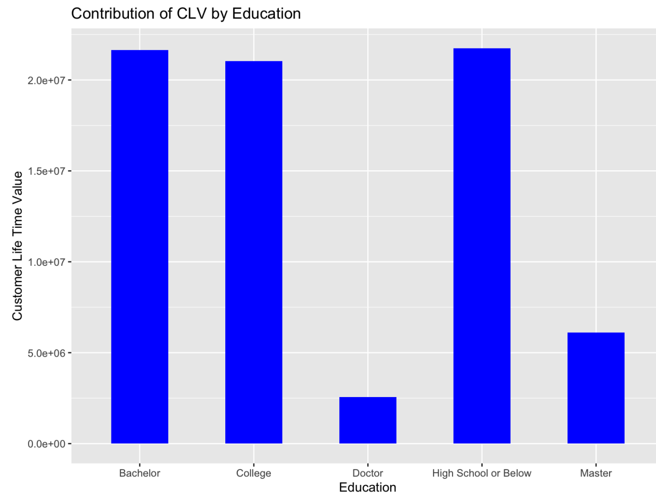
Recommendation:

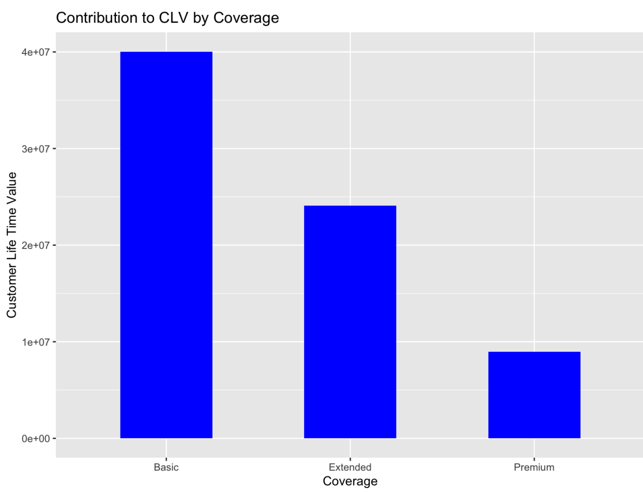
- The Regression Tree Model performs well and provides a clear decision-making structure; however, its performance can be compared to that of the Random Forest's ensemble approach.
- The Random Forest Model continues to stand out as a strong performer, offering a good balance between interpretability and predictive accuracy.

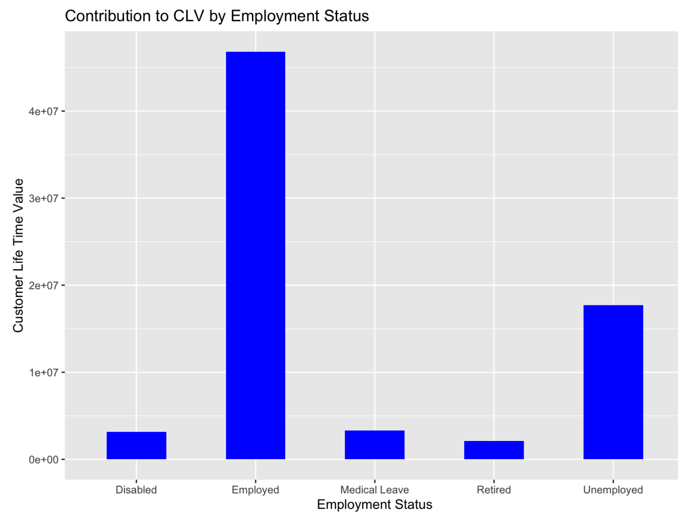
APPENDIX

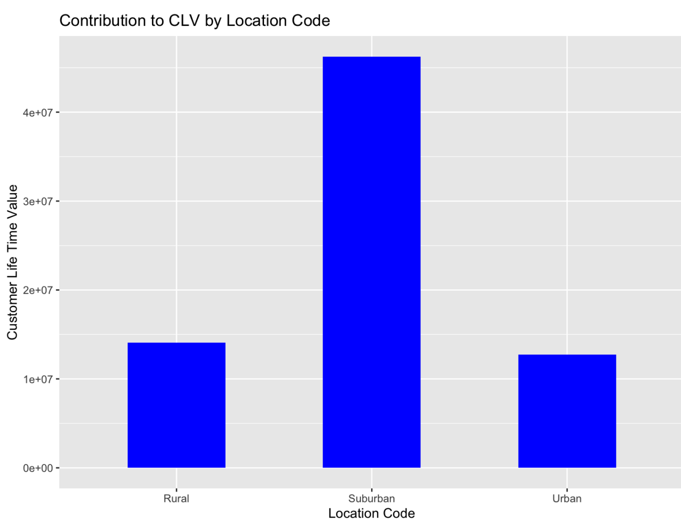
The following are some of the Plots generated from the EDA of the categorical variables.

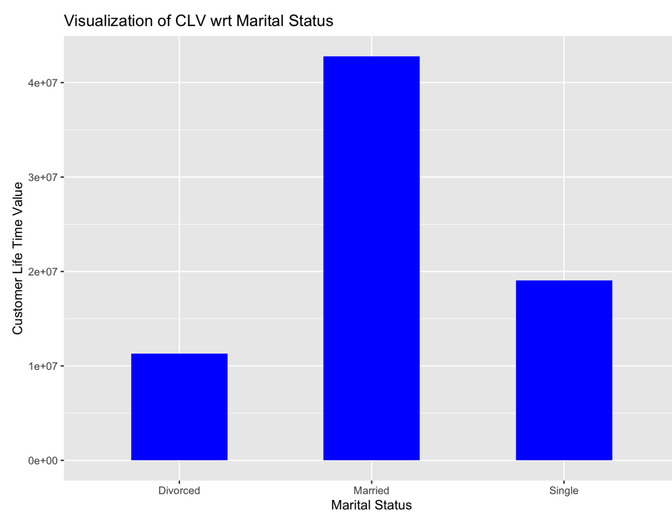


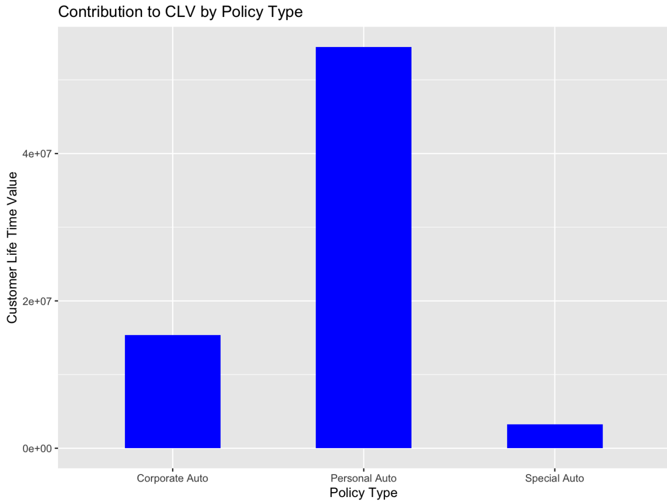


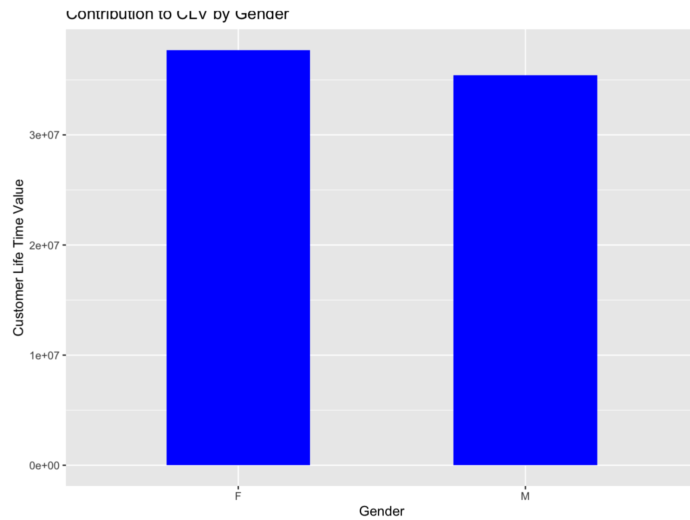


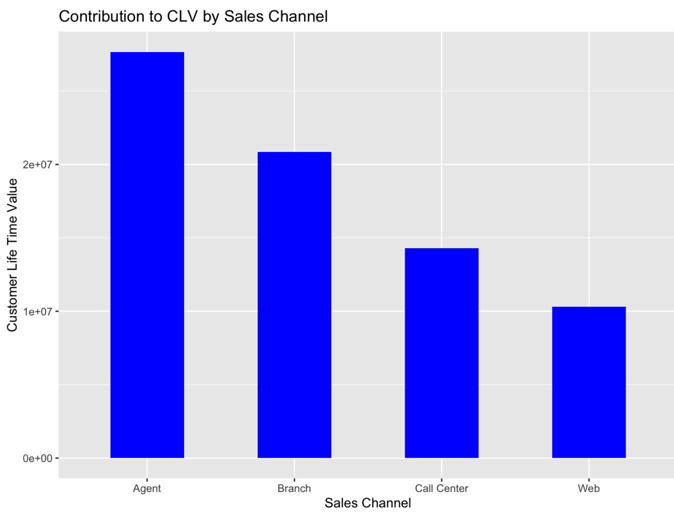


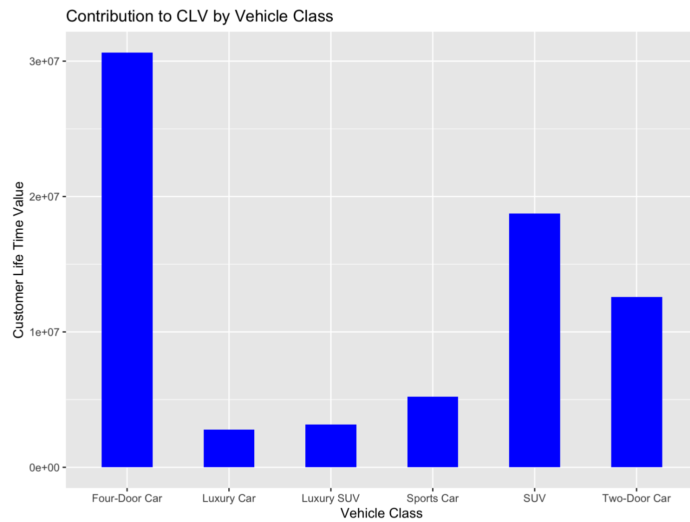












References:

https://github.com/Sarah-2510/Customer-Lifetime-Value-Prediction/blob/main/Code.Rmd