Customer Lifetime Value Prediction Model

Presented By: Abijin Suvedha A

Introduction

In today's competitive market, retaining customers is as important as acquiring new ones. Customer Lifetime Value (CLV) is a key metric that estimates the total revenue a business can expect from a single customer over time. Predicting CLV helps companies identify high-value customers and design effective marketing strategies. This project focuses on developing a machine learning model to predict CLV based on customer purchase behavior and demographic attributes.

Abstract

The objective of this project is to build a predictive model that estimates each customer's lifetime value using available demographic and transactional data. The dataset contains customer details such as state, income, education, policy type, premium amount, vehicle class, and total claim amount. The workflow involves data cleaning, feature engineering, model training, evaluation, and visualization. Regression algorithms like Random Forest and XGBoost are applied to predict the CLV. The model enables targeted marketing by segmenting customers into high, medium, and low-value groups, thereby improving business profitability and customer retention strategies.

Tools Used

- **Programming Language:** Python
- **Libraries:** Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn, XGBoost
- **Environment:** Google Colab
- **Other Tools:** Excel (for initial data inspection and formatting)

Deliverables:

- Python Notebook (.ipynb)
- Trained Model (Pickle File)
- CLV Prediction Output (CSV)
- Visualizations and Report (PDF)

Steps Involved in Building the Project

1. DataCollection:
   Collected customer transaction and profile data containing features like income, vehicle class, premium amount, and claim history.
2. Data Preprocessing:
   - Handled missing values and corrected data types.
   - Encoded categorical variables (e.g., gender, state, education).
   - Normalized numerical features for model stability.
3. Feature Engineering:
   - Derived metrics such as Average Order Value (AOV), Recency, and Frequency.
   - Selected important predictors for CLV such as income, number of policies, and total claim amount.
4. Model Training:
   - Split dataset into training and testing sets (80:20).
   - Trained Random Forest Regressor and XGBoost Regressor models.
   - Tuned hyperparameters using Grid Search for optimal performance.
5. Model Evaluation:
   - Evaluated using Mean Absolute Error (MAE) and Root Mean Square Error (RMSE).
   - Compared models to select the best performer for prediction.
6. Customer Segmentation:
   - Classified customers into High, Medium, and Low CLV segments based on predicted values.
   - Visualized results using histograms and box plots for insights.

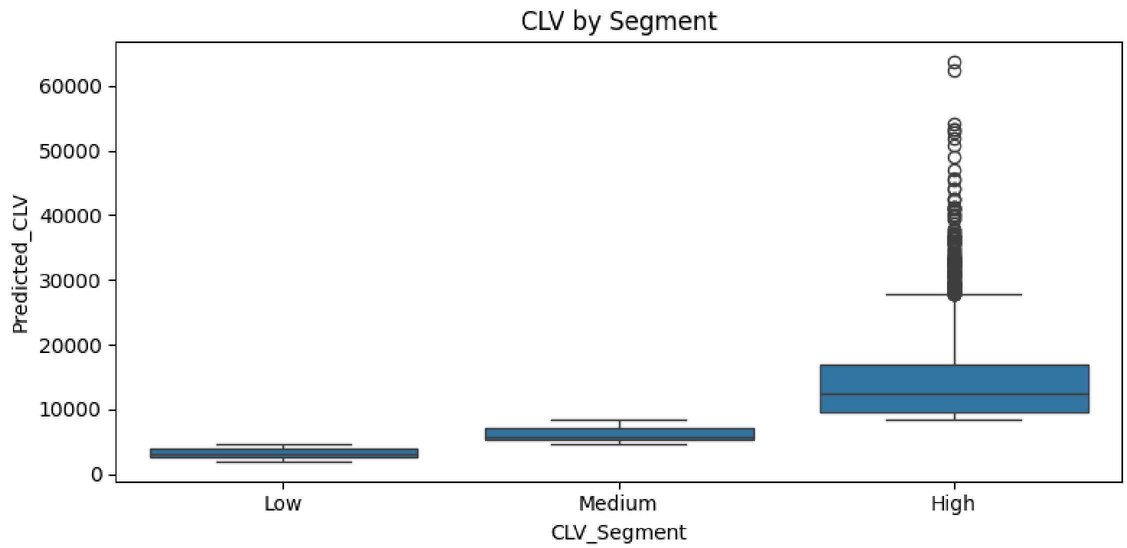
Output:

```
Loaded data with shape: (9134, 24)

Columns:
['Customer', 'State', 'Customer Lifetime Value', 'Response', 'Coverage',
'Education', 'Effective To Date', 'EmploymentStatus', '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', 'Policy', 'Renew Offer Type', 'Sales
Channel', 'Total Claim Amount', 'Vehicle Class', 'Vehicle Size']
```

```
Sample rows:
  Customer       State  Customer Lifetime Value Response  Coverage
Education  \
0  BU79786  Washington              2763.519279       No     Basic
Bachelor
1  QZ44356     Arizona              6979.535903       No  Extended
Bachelor
2  AI49188      Nevada             12887.431650       No   Premium
Bachelor
3  WW63253  California              7645.861827       No     Basic
Bachelor
4  HB64268  Washington              2813.692575       No     Basic
Bachelor

  Effective To Date EmploymentStatus Gender  Income  ...  \
0           2/24/11         Employed      F   56274  ...
1           1/31/11       Unemployed      F       0  ...
2           2/19/11         Employed      F   48767  ...
3           1/20/11       Unemployed      M       0  ...
4            2/3/11         Employed      M   43836  ...

  Months Since Policy Inception Number of Open Complaints  Number of
Policies  \
0                             5                         0
1
1                            42                         0
8
2                            38                         0
2
3                            65                         0
7
4                            44                         0
1

      Policy Type        Policy  Renew Offer Type  Sales Channel  \
0  Corporate Auto  Corporate L3            Offer1          Agent
1   Personal Auto   Personal L3            Offer3          Agent
2   Personal Auto   Personal L3            Offer1          Agent
3  Corporate Auto  Corporate L2            Offer1    Call Center
4   Personal Auto   Personal L1            Offer1          Agent

  Total Claim Amount  Vehicle Class Vehicle Size
0         384.811147   Two-Door Car      Medsize
1        1131.464935  Four-Door Car      Medsize
2         566.472247   Two-Door Car      Medsize
3         529.881344            SUV      Medsize
4         138.130879  Four-Door Car      Medsize

[5 rows x 24 columns]
Using target column: Customer Lifetime Value
Dropped 0 rows with missing target
Numeric cols used: ['Income', 'Monthly Premium Auto', 'Months Since Last
Claim', 'Months Since Policy Inception', 'Number of Open Complaints',
'Number of Policies', 'Total Claim Amount']
Categorical cols used (sample): ['State', 'Response', 'Coverage',
'Education', 'EmploymentStatus', 'Gender', 'Location Code', 'Marital
Status', 'Policy Type', 'Policy']
Final categorical columns encoded: ['Response', 'Gender', 'Coverage',
'Location Code', 'Marital Status', 'Policy Type', 'Vehicle Size', 'Renew
Offer Type', 'Sales Channel', 'State', 'Education', 'EmploymentStatus']
```

```
Train shape: (7307, 20) Test shape: (1827, 20)
Fitting 3 folds for each of 12 candidates, totalling 36 fits
/tmp/ipython-input-196438555.py:42: UserWarning: Could not infer format, so
each element will be parsed individually, falling back to `dateutil`. To
ensure parsing is consistent and as-expected, please specify a format.
  df['Effective To Date'] = pd.to_datetime(df['Effective To Date'],
errors='coerce')
Random Forest best params: {'regressor__max_depth': 20,
'regressor__min_samples_leaf': 1, 'regressor__n_estimators': 200}
Random Forest MAE: 1460.510, RMSE: 3973.858, R2: 0.694
Fitting 3 folds for each of 8 candidates, totalling 24 fits
XGBoost best params: {'regressor__learning_rate': 0.05,
'regressor__max_depth': 6, 'regressor__n_estimators': 100}
XGBoost MAE: 1585.078, RMSE: 4090.562, R2: 0.675
Best model: RandomForest with MAE 1460.510
Saved model to models/best_clv_model.joblib
Saved predictions to outputs/clv_predictions.csv
Saved segmented predictions to outputs/clv_segments.csv
Saved plots to outputs/
Done.
```

clv_by_segment.png × | clv_predictions.csv | clv_segments.csv | predicted_clv_hist.png



clv_by_segment.png | clv_predictions.csv × | clv_segments.csv | predicted_clv_hist.png

1 to 25 of 9134 entries Filter

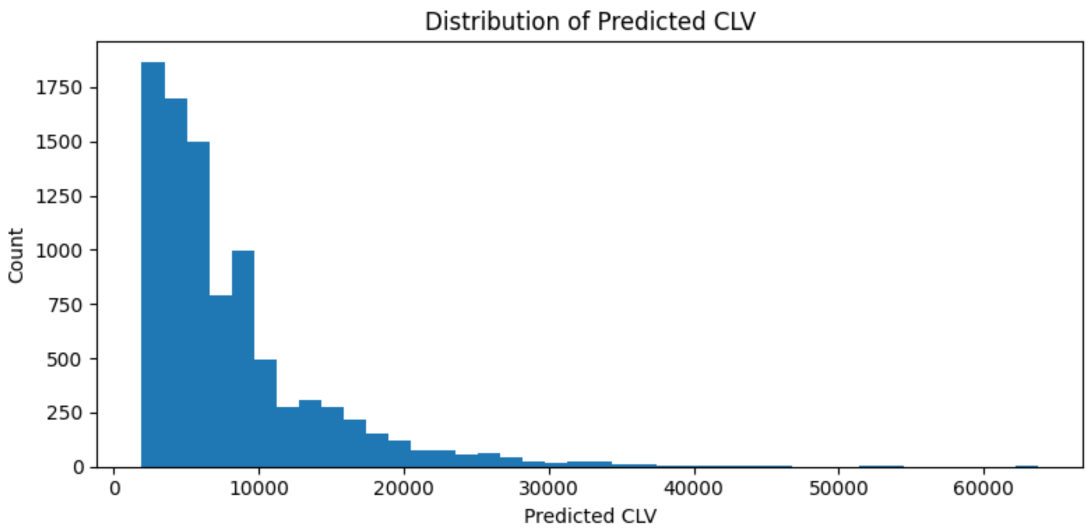
Customer	State	Customer Lifetime Value	Response	Coverage	Education	Effective To Date	EmploymentStatus	Gender	Income	Location Code	Marital Status	Monthly Premium Auto	Months Since Last Claim	Months Since P
BU79786	Washington	2763.519279	No	Basic	Bachelor	2011-02-24	Employed	F	56274	Suburban	Married	69	32	5
QZ44356	Arizona	6979.535903	No	Extended	Bachelor	2011-01-31	Unemployed	F	0	Suburban	Single	94	13	42
AI49188	Nevada	12887.43165	No	Premium	Bachelor	2011-02-19	Employed	F	48767	Suburban	Married	108	18	38
WW63253	California	7645.861827	No	Basic	Bachelor	2011-01-20	Unemployed	M	0	Suburban	Married	106	18	65
HB64268	Washington	2813.692575	No	Basic	Bachelor	2011-02-03	Employed	M	43836	Rural	Single	73	12	44
OC83172	Oregon	8256.2978	Yes	Basic	Bachelor	2011-01-25	Employed	F	62902	Rural	Married	69	14	94
XZ87318	Oregon	5380.898636	Yes	Basic	College	2011-02-24	Employed	F	55350	Suburban	Married	67	0	13
CF85061	Arizona	7216.100311	No	Premium	Master	2011-01-18	Unemployed	M	0	Urban	Single	101	0	68
DY87989	Oregon	24127.50402	Yes	Basic	Bachelor	2011-01-26	Medical Leave	M	14072	Suburban	Divorced	71	13	3
BQ94931	Oregon	7388.178085	No	Extended	College	2011-02-17	Employed	F	28812	Urban	Married	93	17	7
SX51350	California	4738.992022	No	Basic	College	2011-02-21	Unemployed	M	0	Suburban	Single	67	23	5
VQ65197	California	8197.197078	No	Basic	College	2011-01-06	Unemployed	F	0	Suburban	Married	110	27	87
DP39365	California	8798.797003	No	Premium	Master	2011-02-06	Employed	M	77026	Urban	Married	110	9	82
SJ95423	Arizona	8819.018934	Yes	Basic	High School or Below	2011-01-10	Employed	M	99845	Suburban	Married	110	23	25
IL66569	California	5384.431665	No	Basic	College	2011-01-18	Employed	M	83689	Urban	Single	70	21	10
BW63560	Oregon	7463.139377	No	Basic	Bachelor	2011-01-17	Employed	F	24599	Rural	Married	64	12	50

clv_by_segment.png clv_predictions.csv clv_segments.csv predicted_clv_hist.png

1 to 25 of 9134 entries Filter

Customer	State	Customer Lifetime Value	Response	Coverage	Education	Effective To Date	EmploymentStatus	Gender	Income	Location Code	Marital Status	Monthly Premium Auto	Months Since Last Claim	Months Since P
BU79786	Washington	2763.519279	No	Basic	Bachelor	2011-02-24	Employed	F	56274	Suburban	Married	69	32	5
QZ44356	Arizona	6979.535903	No	Extended	Bachelor	2011-01-31	Unemployed	F	0	Suburban	Single	94	13	42
AI49188	Nevada	12887.43165	No	Premium	Bachelor	2011-02-19	Employed	F	48767	Suburban	Married	108	18	38
WW63253	California	7645.861827	No	Basic	Bachelor	2011-01-20	Unemployed	M	0	Suburban	Married	106	18	65
HB64268	Washington	2813.692575	No	Basic	Bachelor	2011-02-03	Employed	M	43836	Rural	Single	73	12	44
OC83172	Oregon	8256.2978	Yes	Basic	Bachelor	2011-01-25	Employed	F	62902	Rural	Married	69	14	94
XZ87318	Oregon	5380.898636	Yes	Basic	College	2011-02-24	Employed	F	55350	Suburban	Married	67	0	13
CF85061	Arizona	7216.100311	No	Premium	Master	2011-01-18	Unemployed	M	0	Urban	Single	101	0	68
DY87989	Oregon	24127.50402	Yes	Basic	Bachelor	2011-01-26	Medical Leave	M	14072	Suburban	Divorced	71	13	3
BQ94931	Oregon	7388.178085	No	Extended	College	2011-02-17	Employed	F	28812	Urban	Married	93	17	7
SX51350	California	4738.992022	No	Basic	College	2011-02-21	Unemployed	M	0	Suburban	Single	67	23	5
VQ65197	California	8197.197078	No	Basic	College	2011-01-06	Unemployed	F	0	Suburban	Married	110	27	87
DP39365	California	8798.797003	No	Premium	Master	2011-02-06	Employed	M	77026	Urban	Married	110	9	82
SJ95423	Arizona	8819.018934	Yes	Basic	High School or Below	2011-01-10	Employed	M	99845	Suburban	Married	110	23	25
IL66569	California	5384.431665	No	Basic	College	2011-01-18	Employed	M	83689	Urban	Single	70	21	10
BW63560	Oregon	7463.139377	No	Basic	Bachelor	2011-01-17	Employed	F	24599	Rural	Married	64	12	50

clv_by_segment.png clv_predictions.csv clv_segments.csv predicted_clv_hist.png



Conclusion

The Customer Lifetime Value Prediction Model provides a data-driven approach for customer segmentation and marketing optimization. By leveraging machine learning techniques, businesses can prioritize valuable customers, personalize offers, and improve retention. The project demonstrates the integration of analytics and business strategy to enhance customer relationship management.