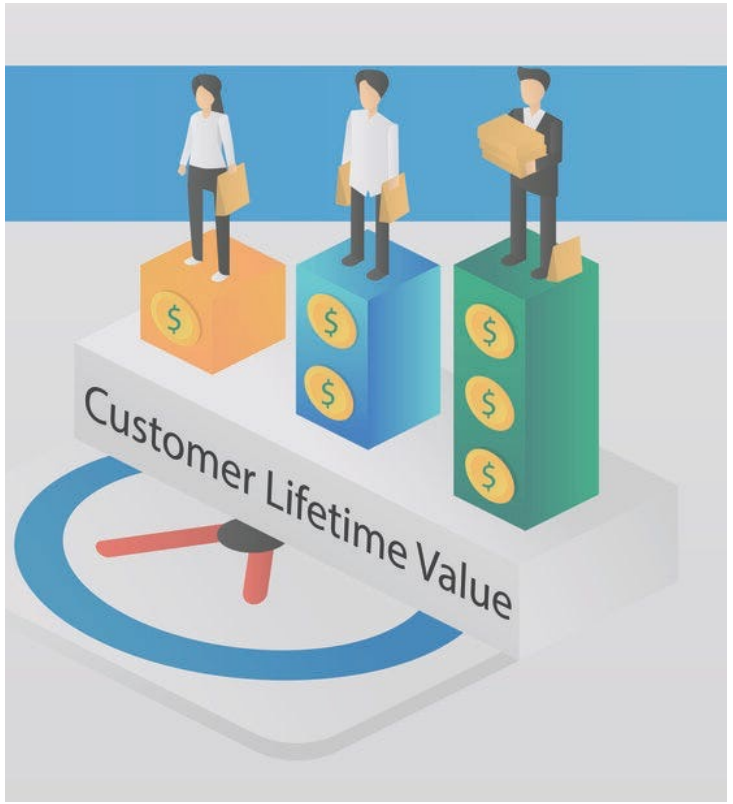
Customer Lifetime Value Prediction Using Regression and Probabilistic Models



Introduction:

- Customer lifetime value (CLV or CLTV) is a measure of the overall net profit a business may anticipate from a customer throughout the course of their relationship. It considers the customer's first purchase, subsequent transactions, and the typical length of their connection with the business.
- This project focuses on predicting Customer Lifetime Value (CLV), a critical metric for businesses, using a combination of Regression and Probabilistic Models.
- We begin with data preprocessing and employ Linear Regression to build an initial CLV model, identifying key factors influencing customer value.
- We add probabilistic models to account for the inherent uncertainty in consumer spending, which enables us to depict CLV as a distribution for more accurate predictions.[1]

Applications:

1. Customer Segmentation: Based on their estimated lifetime value, CLV is utilized to classify customers into different groups. This helps the business to improve their marketing strategies.

2. Customer Acquisition: Predictions of CLV help to locate possible high-value clients. Businesses can focus their acquisition efforts on people or groups that are more likely to become loyal, high-value clients in the long run.

3. Churn Prediction: Customer churn can be predicted using CLV models. Businesses can take proactive actions to retain consumers by identifying those whose CLV is dropping and offering them special promotions.[2]

Motivation and Problem Statement:

- CLV prediction enables data-driven decision-making by providing insights into customer behaviours and preferences. It helps businesses allocate resources more efficiently, optimizing marketing and retention efforts.
- Focusing on CLV enhances long-term profitability by nurturing high-value customer relationships. CLV empowers businesses to offer personalized experiences and gain a competitive edge in customer satisfaction and loyalty.
- This project aims to improve CLV estimation by integrating Regression and Probabilistic Models to account for uncertainty and variability in customer spending, thereby enabling more informed marketing and resource allocation strategies. Also estimates the CLV of the new customer given the previous transaction data.

Literature Survey :

- Yuechi Sun et al. used BG/NBD and Gamma-Gamma models,RFM Analysis to predict the CLV and the probability of a customer being alive of an Online-Retail Store.
- The model is trained with the data of 25 months and the next 12 months transactions are predicted.
- For Evaluating the performance, they divided the data into Calibration and holdout and plotted a graph between the Predicted purchases and the Actual Purchases.[4]

- Rajeev Gupta et.al used Linear Regression, Decision Tree Regression, Random Forest Regression, AutoML Regression models to estimate the CLV of the Insurance Company.
- The data set contains the information of 9134 different nt

Model	RMSE	MAE	R2 Score
Linear Regression	0.573	0.443	0.263
Decision Tree Regression	0.212	0.100	0.898
Random Forest Regression	0.211	0.103	0.899
Auto ML Regression	0.210	0.104	0.901

- Results:

Literature Survey :

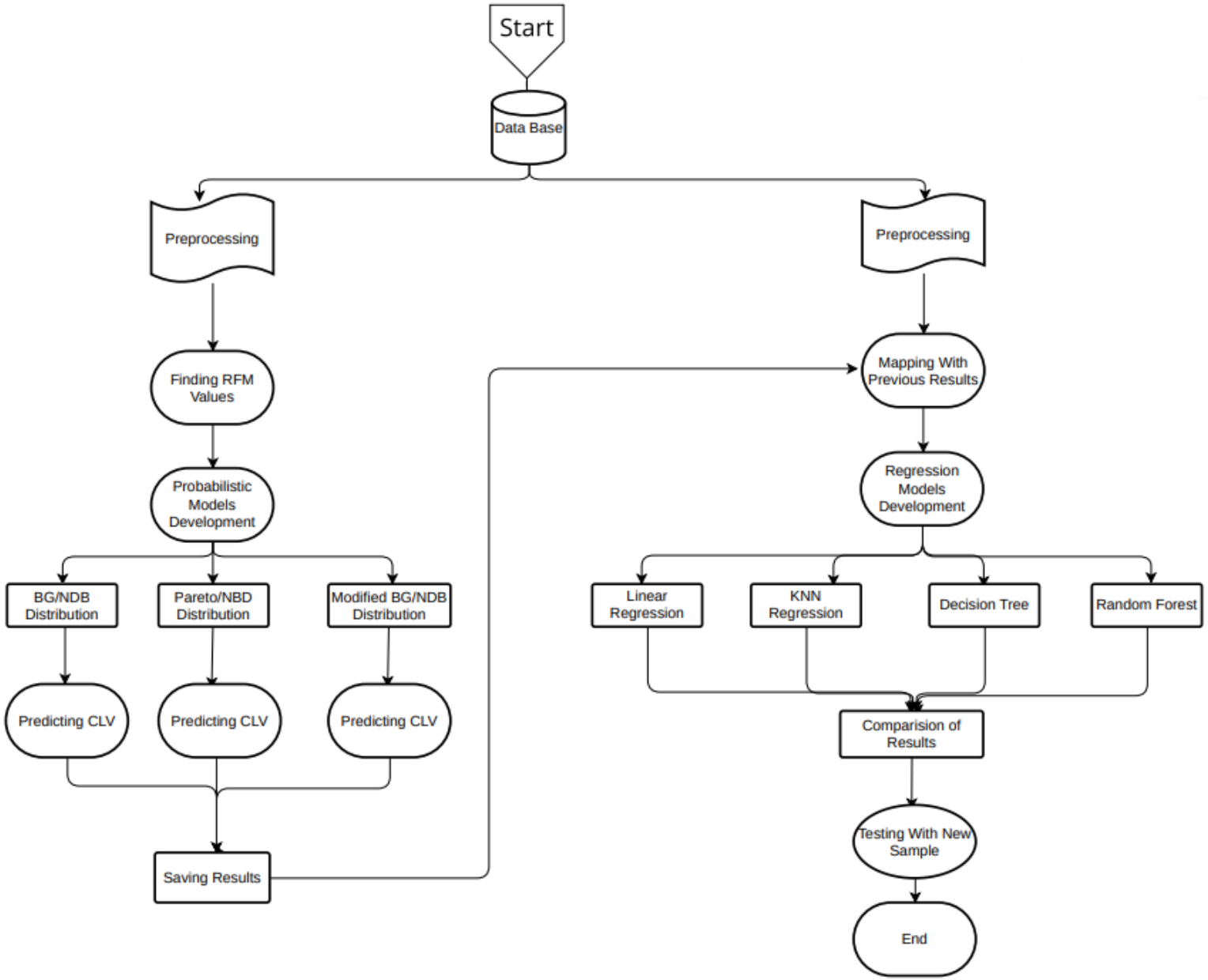
- Ravi Kant et.al focuses on the evaluation of machine learning algorithms for predicting CLV. The study evaluates and compares the performance of four popular machine learning algorithms linear regression, decision tree, support vector machine, and random forest.
- Dataset contains 9134 order line rows, with a total of 25 features.[6]
- Results:

Model	RMSE	R Squared
Linear Regression	0.8702	0.2273
Decision Tree	0.4070	0.8309
Random Forest	0.3064	0.9042
Support Vector Machine	0.9289	0.1197

- A Segmentation method using Fuzzy-AHP and RFM analysis is suggested by Anu Gupta Aggarwal et al. They Performed this on an open dataset from Kaggle. For clustering they used k-mean clustering algorithm.
- Customers are divided into 8 Cluster based on their Recent transactions, Frequency of the purchases and their Monetary value.
- According to their study Monetary (0.603),follow Frequency (0.321) and the least Recency (0.076). Based on these weights, ranked the customer segmentations. [7]

Literature Survey - Base Paper:

- Asiman Mammadzada et al. suggested an approach to find the CLV using BG/NBD, Gamma-Gamma Distribution and used RFM analysis to predict the customer value of a bank.
- The model is trained with the data of 6 months and the next 15 days transactions are predicted.
- For Evaluating the performance before Implementation, they divided the data into Calibration and holdout and plotted a graph between the Predicted purchases and the Actual Purchases.
- And after Implementation of the model the accuracy obtained was 92.1%.[9]


Start
Data Base
Preprocessing
Finding RFM Values
Probabilistic Models Development
BG/NDB Distribution
Pareto/NBD Distribution
Modified BG/NDB Distribution
Predicting CLV
Predicting CLV
Predicting CLV
Saving Results
Preprocessing
Mapping With Previous Results
Regression Models Development
Linear Regression
KNN Regression
Decision Tree
Random Forest
Comparision of Results
Testing With New Sample
End

Flow Chart:

Dataset

- online_retail_II (cltv data set).csv
- This is a Data set of a Company consists of their sales in the year 2009-2010 and 2010-2011
- This data set contains 8 attributes (Invoice No, Stock Code, Description, Quantity, Date of Purchase, Price, Customer ID, Country).
- Total Samples in the dataset are 1067371.
- There are 4185 unique Customers.

Preprocessing:

- Null Values from the dataset has been removed.
- Grouped according to the Customer Id.
- Most of the customers are from the United Kingdom followed by the Germany, EIRE & France.
- Most of the transaction happened in the month of November.
- Year-2010 has most no.of transaction followed by 2011 and 2009.
- There are 4185 unique Customers.
- In second preprocessing the dataset has transformed into the Customer ID, 25 months of the transaction Value and mapped to the CLV predicted using the Probabilistic Models.

Model Development:

- The CLV has been found using probabilistic models.
- BG/NBD, Pareto/NBD, and Modified BG/NBD models are used for finding the CLV.
- Using the Calibration and Holdout method the performance of the Probabilistic models has been measured.
- K-mean Clustering model and RFM values are used to segment the customers.
- Then the Customer ID, Monthly transaction of 26 months, and the predicted CLV are given as the Inputs to the Regression Models(Linear Regression, KNN Regression, SVM Regressor, Decision Tree Regression, Random Forest Regression) are used to train the model.
- Accuracy, Root Mean Square Error, and R-squared error of each Model has been used to Evaluate the Performance of the Model.

BG/NBD Distribution:

The BG/NBD model is a probabilistic integrated model that explains the purchasing and churning behaviors of consumers.

$$E(Y(t) \mid X = x, t_x, T, r, \alpha, a, b) = \frac{\frac{a+b+x-1}{a-1}\left[1 - \left(\frac{\alpha+T}{\alpha+T+t}\right)^{r+x} {}_2F_1\left(r+x, b+x; a+b+x-1; \frac{t}{\alpha+T+t}\right)\right]}{1 + \delta_{x>0}\frac{a}{b+x-1}\left(\frac{\alpha+T}{\alpha+t_x}\right)^{r+x}}$$

x → Frequency of customers who have made at least two purchases
tx → Customer's recency value (must be calculated individually for each customer)
T →Time since the customer's first purchase. Age of customer for company. Tenure.
r, alfa → Difference in transaction rate between customers parameters of gamma distribution
a,b → Beta distribution parameters expressing drop rate

Pareto/NBD Distribution:

The Pareto/NBD model differs in that a consumer can only leave immediately following a transaction. This greatly streamlines calculations, but it has the disadvantage that a consumer cannot leave until a transaction is completed. According to the Pareto/NBD model, a consumer may leave at any time.

$$E(T_x) = \frac{r}{\alpha+s} \times \left(s + T + \frac{1}{\beta} - \left(s + t_x + \frac{1}{\beta} \right) \times e^{-(\alpha+\beta)\times T} \times {}_1F_1\left(\alpha+\beta, \alpha+\beta; (\alpha+\beta) \times (T+t_x) \right) \right)$$

$E(Tx)$ = Expected number of transactions for a customer.
r = Transaction rate parameter.
α and β = Parameters of the Pareto/NBD model.
s = Scale parameter.
$T\,x$ = Customer's recency value.
T = Time since the customer's first purchase (tenure).
$1F1$ = Kummer's confluent hypergeometric function.

BG/NBD Distribution:

Coefficient Obtained while fitting the model to dataset are:
a: 0.15, alpha: 49.94, b: 2.11, r: 0.67

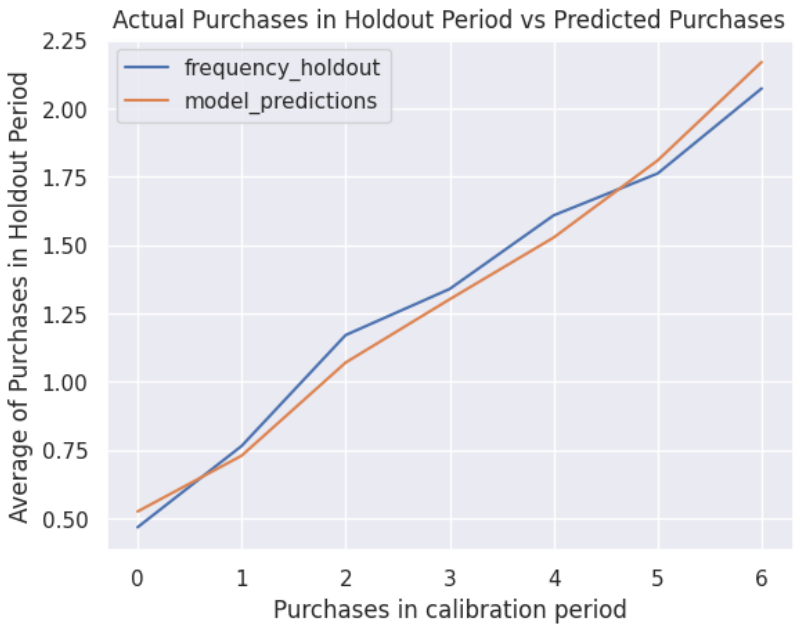
Pareto/NBD Distribution:

Coefficient Obtained while fitting the model to dataset are:
alpha: 63.88, beta: 124.23, r: 0.83, s: 0.16

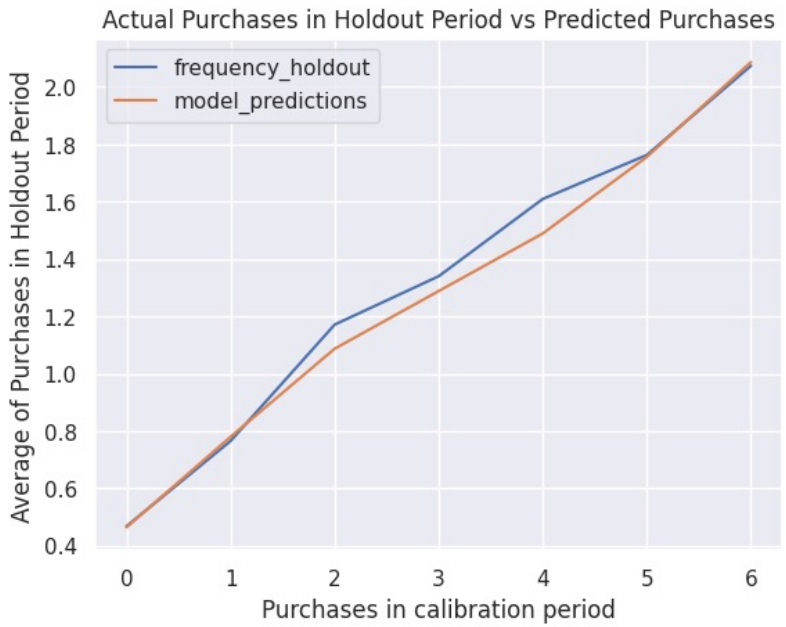
Modified BG/NBD Distribution:

Coefficient Obtained while fitting the model to dataset are:
a: 0.18, alpha: 57.90, b: 2.05, r: 0.8

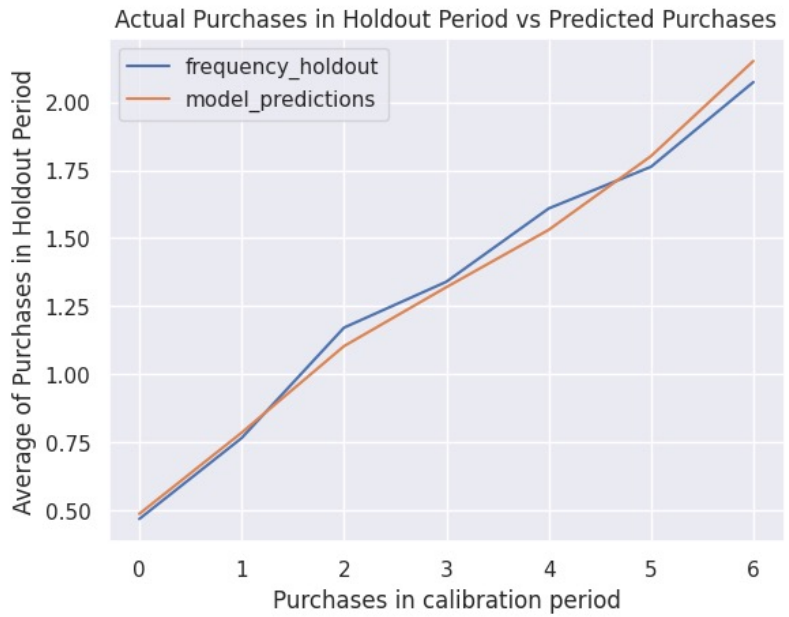
Probabilistic Models Performance:



BG/NBD Distribution



Pareto/NBD Distribution



Modified BG/NBD Distribution

Probabilistic Models Performance:

	BG-NBD	Pareto-NBD	MBG-NBD
MSE Purchase Error	4.337883	4.335935	4.346083
RMSE Purchase Error	2.082758	2.082291	2.084726
Avg Purchase Error	0.411798	0.412367	0.417090

Results:

CLV:

	Customer ID	BG-NBD	Pareto-NBD	MBG-NBD
0	12347.0	7110.166375	6427.083703	7003.290033
1	12348.0	2512.081673	2243.170549	2522.287420
2	12349.0	3449.798774	3184.476671	3504.589499
3	12352.0	3100.010138	2760.845696	3031.895454
4	12353.0	536.035114	474.570351	585.402211

Results:

Segmentation:

	Customer ID	frequency	recency	T	monetary_value	predicted_purchases	actual_30	Error	Expected_Avg_Sales	predicted_clv	CLV	Labels
1	12347.0	7.0	402.0	404.0	717.398571	0.495352	0.522388	0.027036	629.556290	7110.166375	355.508319	Low
2	12348.0	4.0	363.0	438.0	449.310000	0.269178	0.330579	0.061401	409.972301	2512.081673	125.604084	Low
3	12349.0	4.0	717.0	735.0	1107.172500	0.172500	0.167364	-0.005136	842.513995	3449.798774	172.489939	Low
6	12352.0	8.0	356.0	392.0	218.182500	0.566836	0.674157	0.107321	242.209404	3100.010138	155.000507	Low
7	12353.0	1.0	204.0	408.0	89.000000	0.091934	0.147059	0.055125	254.858476	536.035114	26.801756	Low

Algorithms	Accuracy
Linear Regression	93.418005
Random Forest(Estimators=150)	86.573666
Random Forest(Estimators=300)	86.562072
Random Forest(Estimators=250)	86.477140
Random Forest(Estimators=200)	86.407492
Random Forest(Estimators=100)	86.072148
Random Forest(Estimators=50)	86.069422
K-Nearest Neighbors(K=3)	78.822651
K-Nearest Neighbors(K=2)	78.684865
K-Nearest Neighbors(K=1)	78.404233
K-Nearest Neighbors(K=4)	78.195173
K-Nearest Neighbors(K=5)	76.770495
K-Nearest Neighbors(K=6)	75.531666
K-Nearest Neighbors(K=7)	73.891271
K-Nearest Neighbors(K=8)	72.691467
K-Nearest Neighbors(K=9)	71.372031
K-Nearest Neighbors(K=10)	70.148213
Decision Tree	68.688595

Algorithms	Root Mean Square Error
Linear Regression	3341.156856
Random Forest(Estimators=300)	5036.125174
Random Forest(Estimators=100)	5138.202480
Random Forest(Estimators=250)	5138.666398
Random Forest(Estimators=150)	5269.438340
Random Forest(Estimators=200)	5336.495553
Random Forest(Estimators=50)	5659.320183
K-Nearest Neighbors(K=1)	5871.777964
K-Nearest Neighbors(K=2)	6581.648449
K-Nearest Neighbors(K=4)	6989.894130
K-Nearest Neighbors(K=3)	7094.328250
K-Nearest Neighbors(K=5)	7374.047643
K-Nearest Neighbors(K=6)	7378.342657
K-Nearest Neighbors(K=7)	7499.923474
K-Nearest Neighbors(K=8)	7937.210368
K-Nearest Neighbors(K=9)	8176.812549
K-Nearest Neighbors(K=10)	8341.700593
Decision Tree	11514.357774

Regression Models Performance :

Results:

When a New test Sample is given:

Input:a:22251 b:6900 c:18851 d:3869 e:6448 f:3541

new_data = np.array([a,b,c,d,e,f,a,b,c,d,e,f,a,b,c,d,e,f,a,b,c,d,e,f,a])

Output:

The CLV of the Customer using Linear Regression is between $306,836 and $322,761. The average value the CLV is $314,799.

Comparison with Base Paper:

Base Paper:

- Models: BG/NBD, Gamma-Gamma Distribution, RFM Analysis.
- The model is trained with the data of 6 months and the next 15 days transactions are predicted.
- Data set contains the transactional data, the id, date of transaction and the monetary value of 144965 customers.
- For Evaluation Calibration-Holdout method.

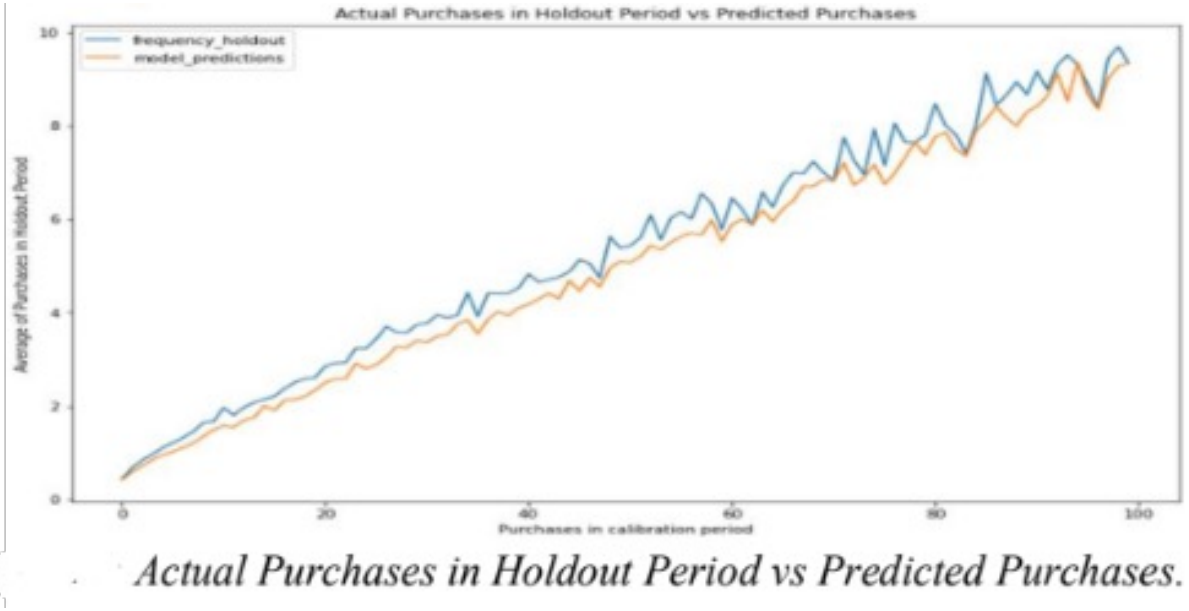
Our Work:

- Models: BG/NBD, Pareto/NBD Distribution, Linear Regression, KNN Regression, Decision Tree, Random Forest, RFM Analysis
- The model is trained with the data of 25 months and the next 6 months transactions are predicted.
- This data set contains 8 attributes (Invoice No, Stock Code, Description, Quantity, Date of Purchase, Price, Customer ID, Country).Total Samples in the dataset are 1067371.There are 4185 unique Customers.
- For Evaluation Calibration-Holdout method, MSE, RMSE.

Comparison with Base Paper:

Base Paper:

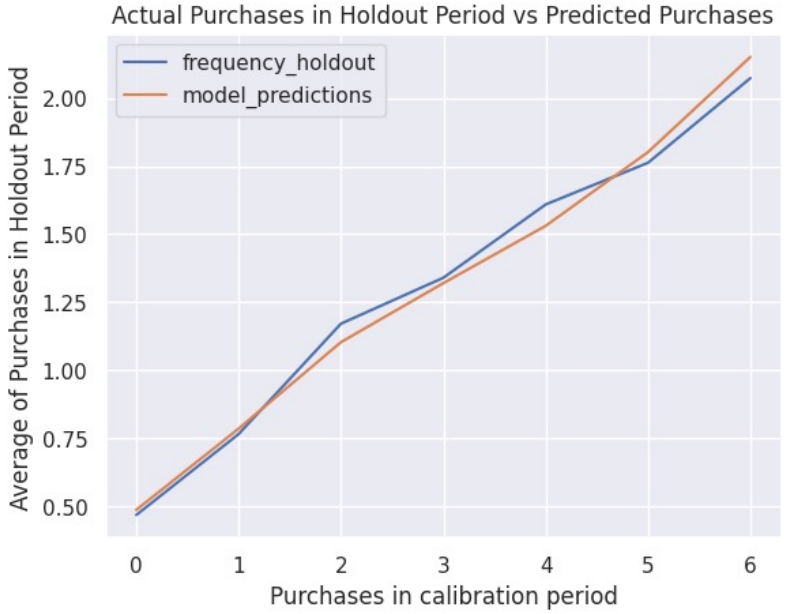
Results:



Actual Purchases in Holdout Period vs Predicted Purchases.

Our Work:

- Results:



Conclusion:

- Predicting Customer Lifetime worth is an essential challenge for companies looking to identify and optimize the worth of their clientele. A variety of methods are available for estimating the future value a customer will generate when using Regression and Probabilistic Models in CLV prediction.
- The work highlights the potential of machine learning in Predicting the CLV particularly using Probabilistic and Regression Models.
- It helps the companies to improve the marketing strategies and concentrate on the High value customers.
- Customer Segmentation divides the customers in Clusters and helps the companies to focus on the Cluster.

Future Scope:

The future scope for Customer Lifetime Value (CLV) prediction using Regression, Probabilistic Models, and RFM (Recency, Frequency, Monetary) analysis is:

1. Advanced Machine Learning Techniques: Future developments may involve the integration of more sophisticated machine learning algorithms.

2. Enhanced Data Integration: Incorporating a wider array of data sources beyond transaction histories and demographics, such as social media interactions, IoT and sentiment analysis, could provide a more comprehensive understanding of customer behavior, further refining CLV predictions.

3. Integration with Customer Experience Management: Integrating CLV predictions with customer experience management platforms will allow businesses to streamline and personalize customer interactions across multiple touchpoints, maximizing CLV.

References:

1. Wikipedia contributors. "Customer lifetime value." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 14 Aug. 2023. Web. 8 Oct. 2023.

2. H. Kailash, K. Kanwar, S. Sonia and R. Kant, "Machine Learning Algorithms for Predicting Customers' Lifetime Value: A Systematic Evaluation," 2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2023, pp. 538-541, doi: 10.1109/ICACITE57410.2023.10182408.

3. Paul D. Berger, Nada I. Nasr Customer lifetime value: Marketing models and applications, Journal of Interactive Marketing, Volume 12, Issue 1,1998

4. Yuechi Sun, Haiyan Liu, Yu Gao,Research on customer lifetime value based on machine learning algorithms and customer relationship management analysis model,Heliyon,Volume 9, Issue 2,2023,e13384,ISSN 2405-8440,https://doi.org/10.1016/j.heliyon.2023.e13384.

5. M. Surti, V. Shah, S. Bharti and R. Gupta, "Customer Lifetime Value Prediction of an Insurance Company using Regression Models," 2023 International Conference for Advancement in Technology (ICONAT), Goa, India, 2023, pp. 1-6, doi: 10.1109/ICONAT57137.2023.10080805.

6. H. Kailash, K. Kanwar, S. Sonia and R. Kant, "Machine Learning Algorithms for Predicting Customers' Lifetime Value: A Systematic Evaluation," 2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), Greater Noida, India, 2023, pp. 538-541, doi: 10.1109/ICACITE57410.2023.10182408

7. Bernar Taşcı, Ammar Omar, Serkan Ayvaz, Remaining useful lifetime prediction for predictive maintenance in manufacturing, Computers & Industrial Engineering, Volume 184,2023,109566,ISSN 0360-8352,https://doi.org/10.1016/j.cie.2023.109566.

8. M. Myburg and S. Berman, "Customer Lifetime Value Prediction with K-means Clustering and XGBoost," 2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Istanbul, Turkey, 2022, pp. 298-302, doi: 10.1109/ASONAM55673.2022.10068602.

9. Mammadzada, E. Alasgarov and A. Mammadov, "Application of BG / NBD and Gamma-Gamma Models to Predict Customer Lifetime Value for Financial Institution," 2021 IEEE 15th International Conference on Application of Information and Communication Technologies (AICT), Baku, Azerbaijan, 2021, pp. 1-6, doi: 10.1109/AICT52784.2021.9620535.

10. S. Maitra, M. Rakib Ahamed, M. Nazrul Islam, M. Abdullah Al Nasim and M. Ashraf, "A Soft Computing Based Customer Lifetime Value Classifier for Digital Retail Businesses," 2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York, NY, USA, 2021, pp. 0074-0083, doi: 10.1109/UEMCON53757.2021.9666546

11. A. Valdivia, "Customer Lifetime Value in Mobile Games: a Note on Stylized Facts and Statistical Challenges," 2021 IEEE Conference on Games (CoG), Copenhagen, Denmark, 2021, pp. 1-5, doi: 10.1109/CoG52621.2021.9619092.

12. A. Tripathi, T. Bagga, S. Sharma and S. Kumar Vishnoi, "Big Data-Driven Marketing enabled Business Performance : A Conceptual Framework of Information, Strategy and Customer Lifetime Value," 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2021, pp. 315-320, doi: 10.1109/Confluence51648.2021.9377156

Thank You