Approach

January 22, 2023

1 Details of Files Present

1. "Data_Exploration.ipynb" => This Notebook explores the dataset
2. "Data_Preprocessing.ipynb" => processing of our data to make it suitable for modelling
3. "Modelling.ipynb" => tested various models here
4. "Alternative_Approach.ipynb" => An alternative approach to our problem
5. "Final_Model.ipynb" => this is the final file and can be used to generate scores similar to what i was able to produce on leaderboard.

2 Objective

The objective was to design a machine learning model to predict the Customer Lifetime Value based on user and policy data.Customer Lifetime Value (CLTV) is a prediction of the total monetary value a customer will bring to a company over their lifetime, for example, the projected revenue from all their purchases.

3 Findings from Exploration of Data

1. Our dataset includes duplicate entries (totaling 26).
2. There are over 12,000 duplicated entries, with varying target variables, which can significantly decrease the model's ability to learn and may indicate the presence of external factors.
3. The distribution of 'CLTV' is heavily skewed to the left, and using 'R-squared' as an accuracy metric may negatively impact the model's performance.
4. The data demonstrates a higher proportion of individuals residing in urban areas, with an income between 5L-10L, holding multiple policies, and predominantly purchasing A policies of the Platinum type.
5. Both the number of policies and the customer's area of residence have a significant impact on their CLTV.
6. There is a notable correlation between claim amount and CLTV.

4 Data Processing

1. I began by eliminating 26 duplicate entries.

2. I applied a feature engineering function to improve the suitability of the income and marital status columns for analysis.
3. I scaled numerical columns to prevent disproportionate influence on the models.
4. I employed the Pandas "dummy" function to one-hot encode the data.
5. Despite initial efforts, feature engineering did not yield significant improvements, so I abandoned the approach.

5 Modelling

1. Different models were evaluated for their ability to accurately predict Customer Lifetime Value.
2. The CatBoost and XGBRF models demonstrated improved performance compared to other models, but there was little distinction between them.
3. Attempts to fine-tune model parameters did not result in any notable improvements.
4. Stacking of various models was able to improve performance marginally.

6 Final Model

1. The CatBoost model was chosen for its superior accuracy and interpretability. Stacking models were not selected as interpretability was prioritized.
2. The model's interpretability is further improved by highlighting the importance of different features.
3. The CatBoost model required minimal preprocessing and additional feature engineering did not significantly enhance performance, therefore it was not implemented.
4. The final model can be found in the "Final_Model.ipynb" file.

7 Suggestions

1. The presence of duplicate features suggests that other important variables may be impacting the results. One suggestion would be to incorporate new features such as age, employment type, and policy pricing, as these could significantly enhance the model's predictive abilities.
2. Additionally, it may be beneficial to categorize 'CLTV' into ranges. This approach is detailed in the notebook titled "Alternative_Approach"

8 Thanks

1. Github Repository link- https://github.com/hemant1456/Analytics-Vidhya-Job-A-Thon-Jan-2023
2. My LinkedIn profile - https://www.linkedin.com/in/hemantbhambhu/
3. Sometimes I write on 'Medium' about data science and other things- https://medium.com/@hemantbhambhu