Customer Lifetime Value (LTV) Prediction

- Project Summary

Introduction

Understanding how much revenue a customer is likely to bring over the course of their relationship with a company is vital for business growth. This metric, known as Customer Lifetime Value (LTV), enables companies to make informed decisions regarding customer acquisition, retention, and resource allocation. The objective of this project is to develop a machine learning model to predict LTV based on customer behavior and demographic attributes, using historical marketing and transactional data.

Abstract

This project demonstrates a practical approach to LTV prediction using Python-based data science techniques. The dataset used includes demographic information and policy data for insurance customers. After preprocessing and engineering relevant features—namely **Recency**, **Frequency**, and **Average Order Value (AOV)**—a regression model is trained to predict the LTV. A Random Forest Regressor was selected for its balance between accuracy and interpretability. The model's performance was evaluated using **MAE** and **RMSE**. Customers were segmented based on predicted LTV into Low, Medium, and High categories. Visualizations and metrics confirmed the model's ability to support data-driven marketing decisions.

Tools and Technologies Used

- **Programming Language**: Python
- **Libraries**:
  - pandas, NumPy – Data manipulation and numerical analysis
  - scikit-learn – Modeling, preprocessing, evaluation
  - matplotlib, seaborn – Visualization
  - xgboost – Boosted decision trees
- **Other**: Microsoft Excel for reviewing/exporting predictions

Project Workflow

1. **Data Understanding and Cleaning**
   a. Read CSV data, parse date columns, and explore basic statistics
   b. Encode categorical variables using label encoding
2. **Feature Engineering**
   a. Recency: Time since the customer's last interaction
   b. Frequency: Number of policies owned

c. AOV: Total Claim Amount divided by Number of Policies

3. **Model Training**
    a. Selected Customer Lifetime Value as the target variable
    b. Trained a Random Forest Regressor using 80-20 train-test split
4. **Model Evaluation**
    a. Evaluated the model with **MAE** and **RMSE**
    b. Plotted feature importance to interpret model predictions
5. **Prediction and Segmentation**
    a. Predicted LTV for each customer
    b. Segmented customers into Low, Medium, and High using quantiles
6. **Visualization and Reporting**
    a. Created charts to show LTV distribution, segment-wise behaviors, and feature correlations
    b. Exported results (Predicted LTV, Segment) to a CSV file

Conclusion

This project showcases how customer behavior and policy information can be leveraged to predict LTV using machine learning. The Random Forest model offered strong performance and interpretability. Customer segmentation based on predicted LTV enables targeted marketing efforts and improved ROI. The methodology developed here can be expanded or integrated into larger customer relationship management (CRM) systems or dashboards. Future work may include using real-time data, incorporating additional behavioral features, or deploying the model into production as an API.