Customer Lifetime Value Prediction System Using the CRISP-DM Methodology

Chandini Saisri Uppuganti
San Jose State University
chandinisaisri.uppuganti@sjsu.edu

October 6, 2024

Abstract

Customer Lifetime Value (CLV) is a critical metric for businesses to assess the value a customer brings throughout their relationship with the company. This paper presents a comprehensive approach for building a CLV prediction system following the CRISP-DM (Cross Industry Standard Process for Data Mining) methodology. Using transaction data, the RFM (Recency, Frequency, Monetary) model, and BG/NBD and Gamma-Gamma modeling techniques, this system accurately predicts the future value of customers. We discuss each phase of CRISP-DM and provide insights into the deployment of the model for real-time predictions.

1 Introduction

Customer Lifetime Value (CLV) is an essential metric for assessing the long-term financial value that a customer brings to a business. Accurate prediction of CLV enables companies to optimize marketing strategies, allocate resources effectively, and prioritize customer retention efforts.

This paper outlines a systematic approach to developing a CLV prediction system using the CRISP-DM methodology. By leveraging transaction data, Recency, Frequency, and Monetary (RFM) values, and applying advanced predictive models, we create a robust system capable of generating accurate CLV predictions.

2 Related Work

There have been multiple studies on CLV prediction using various methodologies. Traditional approaches include RFM analysis and regression models, while more recent work has introduced probabilistic models like BG/NBD and Gamma-Gamma to improve accuracy [**?**]. Our work builds upon these models and follows the CRISP-DM framework, ensuring a structured, reproducible approach to predictive modeling.

3 CRISP-DM Methodology

The CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology consists of six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. This section describes how each phase is applied in the development of our CLV prediction system.

3.1 Business Understanding

The goal of our system is to predict Customer Lifetime Value (CLV) for existing customers. CLV helps businesses identify high-value customers, optimize marketing efforts, and improve customer retention. Understanding these business objectives informs the rest of the data mining process.

3.2 Data Understanding

We utilized transaction data containing customer ID, transaction date, and transaction value. Each transaction contributes to calculating the Recency, Frequency, and Monetary (RFM) values for each customer.

3.3 Data Preparation

The raw transaction data was cleaned and transformed to calculate the RFM values. We used the following method to calculate Recency (days since the last purchase), Frequency (number of purchases), and Monetary (average transaction value).

```
# Calculate Recency, Frequency, and Monetary (RFM)
rfm = df.groupby('customer_id').agg({
    'transaction_date': lambda x: (snapshot_date - x.max()).days,
    'transaction_id': 'count',
    'transaction_value': 'sum'
}).reset_index()

rfm['monetary'] = rfm['transaction_value'] / rfm['transaction_id']
rfm.columns = ['customer_id', 'recency', 'frequency', 'monetary']
```

3.4 Modeling

We employed the BG/NBD and Gamma-Gamma models to predict future customer purchases and their monetary value. The BG/NBD model predicts the number of future transactions a customer will make, while the Gamma-Gamma model predicts the average transaction value.

```
from lifetimes import BetaGeoFitter, GammaGammaFitter

# BG/NBD Model for future purchases
bgf = BetaGeoFitter(penalizer_coef=0.01)
bgf.fit(rfm['frequency'], rfm['recency'], rfm['monetary'])

# Gamma-Gamma Model for monetary value
ggf = GammaGammaFitter(penalizer_coef=0.01)
ggf.fit(rfm['frequency'], rfm['monetary'])

# Predict CLV
rfm['predicted_purchases'] = bgf.conditional_expected_number_of_purchases_up_to_time(6)
rfm['predicted_value'] = ggf.conditional_expected_average_profit(rfm['frequency'], rfm['monetary
rfm['CLV'] = rfm['predicted_purchases'] * rfm['predicted_value']
```

3.5 Evaluation

The model's performance was evaluated using standard metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) by comparing predicted CLV values to actual CLV values over a holdout period.

```
from sklearn.metrics import mean_absolute_error, mean_squared_error

mae = mean_absolute_error(actual_clv, predicted_clv)
rmse = np.sqrt(mean_squared_error(actual_clv, predicted_clv))

print(f"MAE: {mae}, RMSE: {rmse}")
```

3.6 Deployment

We deployed the CLV prediction model using a Flask API to allow real-time CLV predictions based on new customer data. The system accepts input data and returns the predicted CLV, enabling integration with other systems such as CRM platforms.

```
from flask import Flask, request, jsonify

app = Flask(__name__)

@app.route('/predict_clv', methods=['POST'])
def predict_clv():
    data = request.json
    customer_df = pd.DataFrame([data])

    purchases = bgf.conditional_expected_number_of_purchases_up_to_time(6,
                customer_df['frequency'], customer_df['recency'], customer_df['monetary'])
    value = ggf.conditional_expected_average_profit(customer_df['frequency'],
                customer_df['monetary'])
    clv = purchases * value

    return jsonify({"predicted_clv": clv.item()})

if __name__ == '__main__':
    app.run(debug=True)
```

4 Conclusion

By following the CRISP-DM methodology, we developed a CLV prediction system that provides valuable insights into customer behavior and future value. The system utilizes the BG/NBD and Gamma-Gamma models to predict the number of purchases and the monetary value, making it a robust tool for improving marketing strategies and resource allocation.

References

[1] Pyle, D. (1999). *Data Preparation for Data Mining*. San Francisco: Morgan Kaufmann.

[2] Shearer, C. (2000). The CRISP-DM model: The new blueprint for data mining. *Journal of Data Warehousing*, 5(4), 13-22.