Regression model

Credit Card Transaction Prediction

Table of Contents

1. Executive Summary

This project develops a machine learning solution to optimize the timing of credit card reward promotions by predicting when customers are most likely to make transactions in specific categories. Currently, promotional messages are sent at random times, resulting in low engagement rates. By analyzing a comprehensive dataset of 816,859 transactions from 886 unique accounts throughout 2022, this project aims to improve promotional effectiveness through data-driven timing decisions.

The analysis revealed significant patterns in transaction timing, with notable variations across different customer segments and times of day. Transaction volumes show a clear weekly progression, peaking on weekends with approximately 168,000 transactions on Saturdays, compared to about 77,000 on Mondays. Daily patterns indicate highest activity during business hours, with peak volumes around 42,500 transactions per hour between 11 AM and 12 PM.

A Linear Regression model was developed to predict optimal transaction timing, achieving a Mean Absolute Error of 4.12 hours and explaining 35.4% of the variance in transaction timing patterns. This represents a significant improvement over the current random timing approach. The model provides specific timing recommendations for different customer segments, enabling more targeted and effective promotional campaigns.

The implementation of this model requires careful consideration of privacy and ethical implications, particularly given the demographic disparities observed in the data (96.5% urban versus 3.5% rural transactions). Future improvements could include incorporating additional features such as seasonal patterns and geographic factors, as well as implementing more sophisticated machine learning techniques. Despite these limitations, the current model provides a solid foundation for moving from random promotional timing to a data-driven approach, potentially improving customer engagement and promotion effectiveness.

This solution demonstrates the feasibility of using transaction data to optimize marketing timing, while highlighting the importance of balancing technological capabilities with privacy protection and ethical considerations. The recommendations provided by this model can be immediately implemented to improve the bank's promotional strategy, with the framework in place for continued refinement and enhancement.

2. Business Understanding

a. Business Use Case

The bank aims to optimise the timing of its credit card reward promotions by predicting when customers are most likely to make transactions. Currently, promotional messages offering category-specific rewards are sent at random times, leading to low engagement rates. By using predictive models to identify peak transaction hours, the bank can send promotions 1-2 hours before peak transaction times, improving the relevance of the offers and increasing customer engagement.

b. Key Objectives

The objective of the regression model is to optimise the bank's promotion strategy by predicting the best time to send targeted credit card reward offers. Currently, promotions are sent at random times, leading to low engagement. By applying machine learning algorithms to analyse transaction patterns, the bank aims to predict peak transaction hours and send promotions when customers are most likely to engage. This data-driven approach seeks to increase credit card usage, enhance customer loyalty, and boost overall revenue by addressing the inefficiency of the current promotion system. The stakeholders for this initiative are the engaging bank teams and customers with the transactions. Their requirement is to improve the effectiveness of campaigns by sending offers at times when customers are most likely to make purchases. To meet this standard, machine learning algorithms will be applied to predict the optimal timing for sending credit card rewards promotions, thus the project can address the marketing team's goal of increasing customer engagement and driving revenue growth.

3. Data Understanding

a. Basic Data Understanding

The analysis examines banking transaction data from 2022, which was initially distributed across 60 CSV files categorized by age group, gender, and location. The consolidated dataset comprises 816,859 transactions from 886 unique accounts, with each transaction record containing 10 features including credit card number, account number, transaction number, timestamp, category, amount, fraud indicator, and merchant information with geographical coordinates. Analysis reveals significant demographic imbalances: urban transactions dominate with 788,507 transactions (96.5%), split between female (409,841 transactions) and male (378,666 transactions) customers, while rural transactions are notably underrepresented with only 28,352 transactions (3.5%). Several original datasets, particularly in the rural female segment, were empty, indicating potential data collection issues or genuine demographic patterns requiring further investigation. The data quality is robust with no duplicate transactions, and no missing values in critical features, though some merchant latitude and longitude fields contain missing values. There are formatting issues with unix_time and is_fraud: unix_time should be a time-based data type, and is_fraud should be a binary variable, but both are currently shown as object data type. The severe underrepresentation of rural transactions, particularly female rural transactions, should be carefully considered when interpreting the results and drawing conclusions from this analysis.

No	Features	Description	Count	Dtype
1	cc_num	Credit card number	816,859	object
2	acct_num	Account number	816,859	object
3	trans_num	Transaction number	816,859	object
4	unix_time	Time stamp (in seconds)	816,859	object
5	category	Category of transactions	816,859	object
6	amt	Transaction amount ($)	816,859	float64
7	is_fraud	Whether the transaction is fraud or not	816,859	object
8	merchant	Merchant name	816,859	object
9	merch_lat	Merchant latitude	729,820	float64
10	merch_long	Merchant longitude	729,820	float64

Table 1: Basic dataset information

b. Variables of Interest

Trans_date_time

A comprehensive temporal analysis of the transaction data reveals distinct patterns across different time scales. The dataset spans nearly a full year, running from January 6, 2022, to December 31, 2022, encompassing 358 days of transaction records. Weekly transaction patterns show a clear progression, with volume gradually increasing throughout the work week before peaking on weekends. Saturday emerges as the busiest day with approximately 168,763 transactions, followed closely by Friday with 148,563 transactions. In contrast, early weekdays show notably lower activity, with Monday recording the lowest volume at 77,274 transactions.

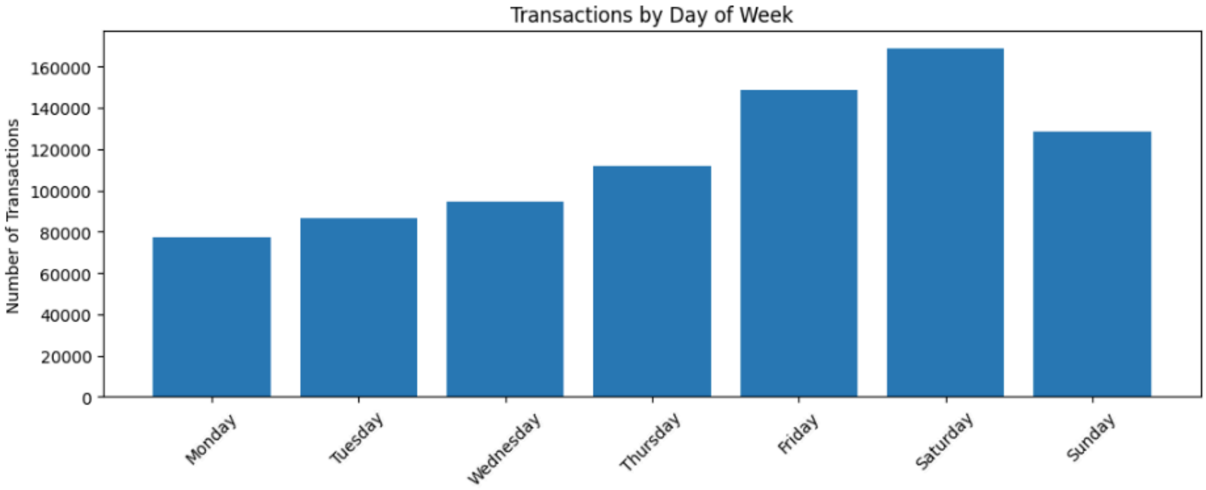


Figure 1: Transactions by Day of Week

Daily transaction patterns exhibit strong temporal characteristics, with the highest activity concentrated during standard business hours. The five busiest hours occur between 9 AM and noon, with peak volumes around 42,500 transactions per hour during the 11 AM to noon period. The hourly distribution follows a distinct pattern: transaction volumes rise sharply from early morning hours, maintain a steady high plateau during business hours (approximately 42,000 transactions per hour between 9 AM and 2 PM), before declining significantly in the afternoon. Evening and overnight hours show consistently lower transaction volumes, maintaining around 26,000-27,000 transactions per hour.

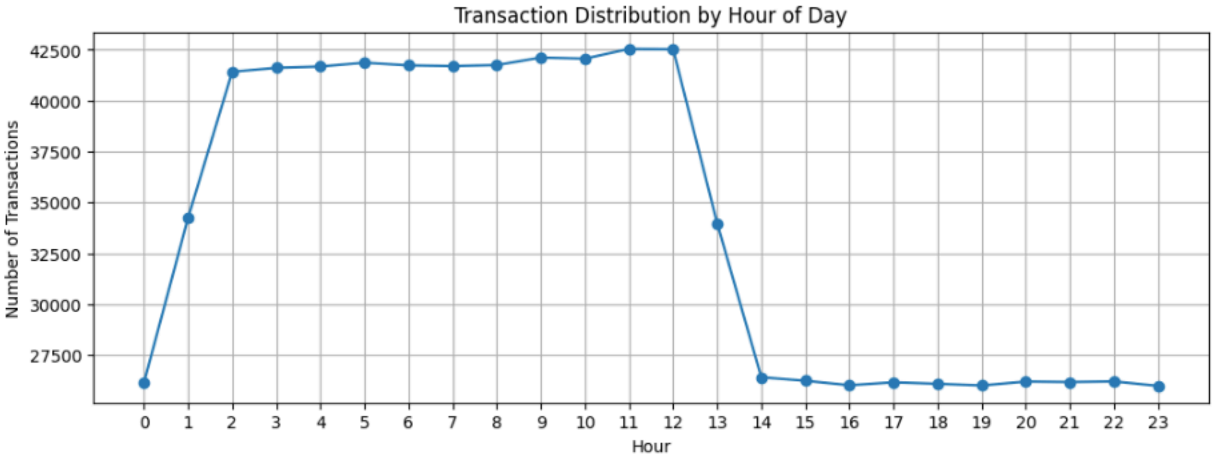


Figure 2: Transactions Distribution by Hour of Day

The monthly analysis shows considerable variation throughout the year, with a notable spike in December (approximately 120,000 transactions), likely corresponding to holiday season shopping patterns. The rest of the year maintains relatively stable monthly volumes, typically ranging between 60,000 to 80,000 transactions per month, with slight variations that could be attributed to seasonal factors or specific promotional periods.

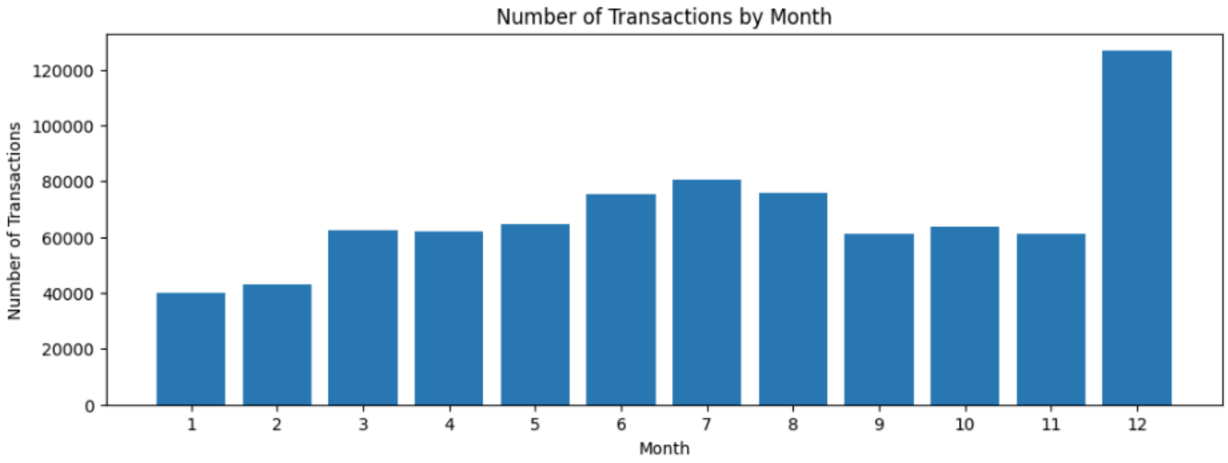


Figure 3: Transactions Distribution by Month

Amount

Analysis of the transaction amounts reveals several distinctive patterns in both the distribution and temporal characteristics of spending. The distribution of transaction amounts is heavily right-skewed, with the majority of transactions concentrated under $100, though the range extends up to $25,000 for some outlier transactions. When examining the log-transformed distribution of amounts, a trimodal pattern emerges, suggesting three distinct transaction clusters that could represent different types of purchasing behaviors.

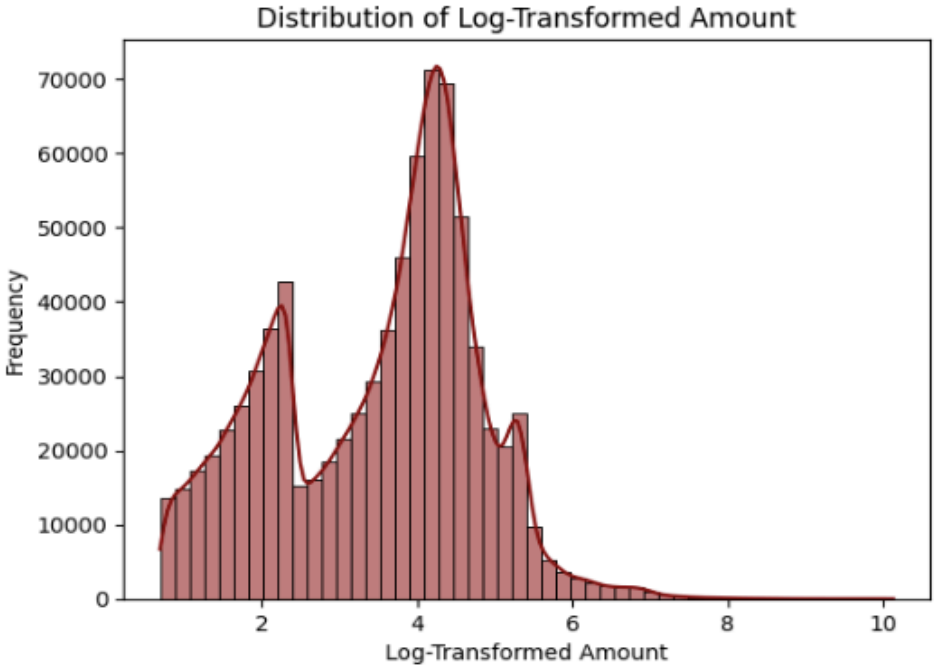


Figure 4: Distribution of Log-Transformed Amount

The hourly analysis of transaction amounts reveals interesting temporal patterns in spending behavior. While the mean transaction amount remains relatively stable throughout the day (ranging from $62 to $79), there is a notable dichotomy between daytime and nighttime median values. During morning hours (2 AM to 1 PM), the median transaction amount hovers around $30-31, but it dramatically shifts to approximately $60-61 during afternoon and evening hours (2 PM onwards). This substantial difference between mean and median values, particularly during morning hours, suggests the presence of high-value outlier transactions that skew the average upward.

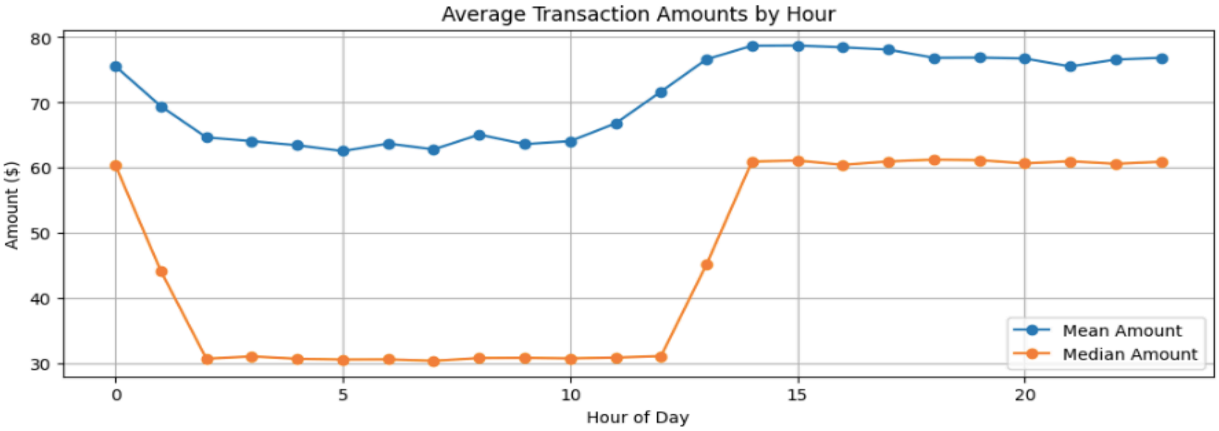


Figure 5: Average Transaction Amounts by Hour

4. Data Preparation

The data preparation phase focused on addressing formatting inconsistencies and establishing the foundational dataset structure. The initial data cleaning process targeted two key formatting issues. First, credit card numbers (cc_num) and account numbers (acct_num) were converted from their original numeric format to string (object) type. This conversion was crucial as these identifiers, while numeric in nature, don't carry any inherent numerical meaning, and treating them as strings prevents unintended numerical interpretations in subsequent analyses. Second, the fraud indicator column (is_fraud) was reformatted to the appropriate integer type to ensure proper handling of this binary flag.

The dataset quality was verified, confirming the absence of duplicate records based on transaction numbers, which ensures data integrity. While the analysis identified missing values in the merchant geographical coordinates (merch_lat and merch_long), these were retained in their original state as they were not critical for the current analysis scope. These geographic data points can be addressed in future iterations if spatial analysis becomes necessary.

Feature engineering in this phase was minimal but foundational, focusing on extracting key demographic and temporal information. Age group, gender, region, and transaction datetime were added as basic features to the dataset. These additions provided essential categorical variables for understanding transaction patterns across different demographic segments. The final prepared dataset contained 15 columns with 816,859 entries, maintaining the full scope of the original transaction data while ensuring proper formatting for subsequent analysis phases.

5. Modeling

This section details the development of a regression model to predict transaction timing, which aims to optimise the delivery of promotional messages by forecasting when customers are likely to make specific types of purchases.

The initial modeling process began with comprehensive feature engineering, where two temporal features - hour and is_weekend - were extracted from the trans_date_time variable to capture potential patterns in transaction timing. For feature selection, SelectKBest with f_regression was employed as it effectively identifies features with strong linear relationships to the target variable, making it particularly suitable for the subsequent linear models. This method was chosen because it provides a statistical basis for feature selection while maintaining interpretability. The preprocessing phase included creating dummy variables for categorical features (category and age_group) and applying StandardScaler to the transaction amount to ensure all features were on a comparable scale. The StandardScale part was taken out of the notebook as the variable was not included in the final model.

The dataset was then split into training (60%), validation (20%), and test (20%) sets, providing sufficient data for model training while maintaining robust validation and testing capabilities. A mean value prediction was selected as the baseline model, which was an appropriate choice as it provides a simple yet meaningful benchmark for regression tasks and helps establish a minimum performance threshold that more complex models should exceed.

The modeling exploration began with three algorithms: Linear Regression, Lasso, and Ridge Regression. The Lasso model underperformed and was eliminated. Linear Regression and Ridge produced identical results, leading to the selection of Linear Regression as the final model due to its simpler architecture and fewer hyperparameters. The decision was further supported by the observation that changing the alpha parameter in Ridge regression had no impact on model performance, suggesting that regularisation was not providing any additional benefit in this case.

The model was fine-tuned by testing different hyperparameter configurations, with fit_intercept=True proving more effective. Through iterative testing, it was discovered that focusing solely on categories and age_groups as features yielded optimal results. While including additional variables provided marginal improvements, the minimal performance gain did not justify the increased model complexity, aligning with the goal of maintaining model simplicity without sacrificing predictive power.

6. Evaluation

a. Evaluation Metrics

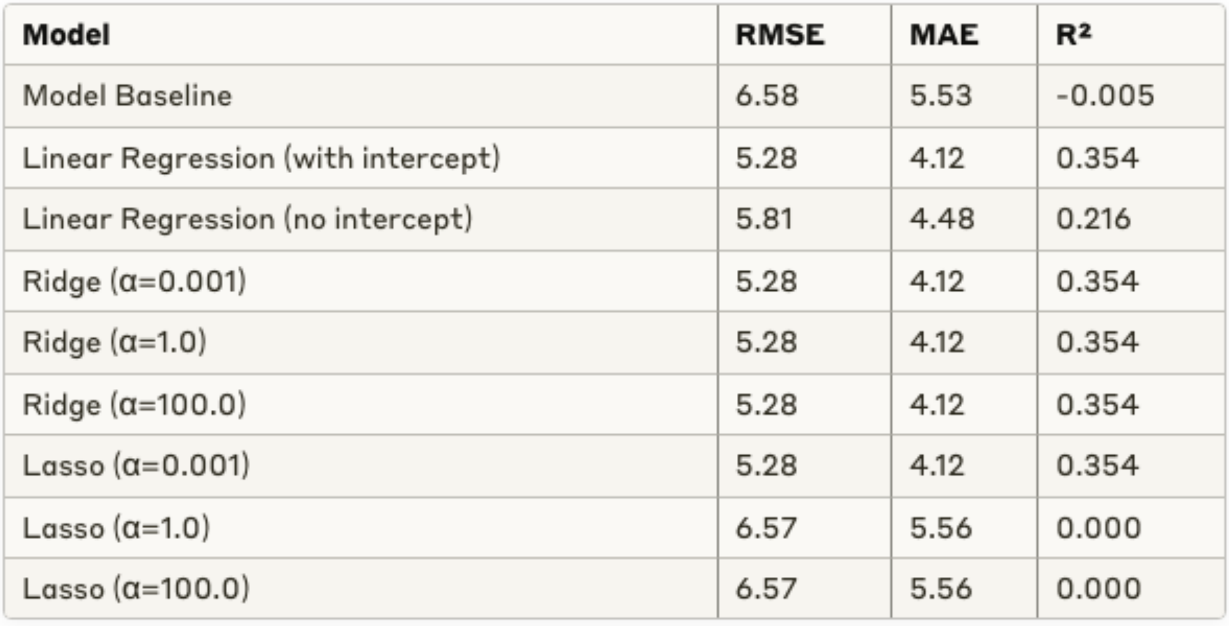
To assess model performance, three complementary evaluation metrics were selected: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2). RMSE and MAE were chosen as primary metrics as they quantify prediction errors in hours, making them directly interpretable in the context of promotional timing optimisation. While both measure prediction accuracy, RMSE penalises larger errors more heavily due to the squared term, providing insight into the model's handling of substantial timing misalignments. MAE offers a more straightforward average of prediction errors, making it particularly useful for communicating model performance to business stakeholders. R^2 was included to measure the proportion of variance explained by the model, providing a standardised measure of goodness-of-fit that facilitates model comparison. These metrics were consistently applied across all models - from the baseline mean value prediction to the various regression approaches (Linear, Ridge, and Lasso) - enabling direct performance comparisons and supporting the final model selection process.

b. Results and Analysis

The experimentation phase tested multiple models and configurations, with Linear Regression (with intercept) and Ridge Regression emerging as the top performers. Both achieved an R^2 score of 0.354, RMSE of 5.28 hours, and MAE of 4.12 hours, representing a substantial improvement over the baseline mean prediction model (R^2 = -0.0053, RMSE = 6.58 hours). Linear Regression without intercept performed notably worse (R^2 = 0.2161), indicating the importance of the intercept term in capturing the underlying patterns in transaction timing.

Key Insights:

- Ridge vs Linear Regression: Ridge regression with varying alpha values (0.001 to 100.0) produced identical results to standard Linear Regression, suggesting minimal multicollinearity among our features and no overfitting issues requiring regularisation.
- Lasso Performance: Lasso regression showed interesting behaviour - with very small alpha (0.001), it matched the performance of Linear Regression, but with larger alpha values (1.0, 100.0), it degraded to baseline performance ($R^2 \approx 0$). This indicates that aggressive feature selection through Lasso regularisation eliminates important predictive signals in our data.
- Error Metrics: The MAE of 4.12 hours for our best models indicates that, on average, predictions are off by about 4 hours, while the higher RMSE (5.28 hours) suggests the presence of some larger errors pulling this metric up.



Model	RMSE	MAE	R²
Model Baseline	6.58	5.53	-0.005
Linear Regression (with intercept)	5.28	4.12	0.354
Linear Regression (no intercept)	5.81	4.48	0.216
Ridge (α=0.001)	5.28	4.12	0.354
Ridge (α=1.0)	5.28	4.12	0.354
Ridge (α=100.0)	5.28	4.12	0.354
Lasso (α=0.001)	5.28	4.12	0.354
Lasso (α=1.0)	6.57	5.56	0.000
Lasso (α=100.0)	6.57	5.56	0.000

Note: All metrics rounded for clarity. RMSE and MAE are in hours.

Figure 5: Regression Model Results

Implications and Areas for Improvement:

The consistent performance across Linear Regression and Ridge suggests we've reached a natural ceiling with our current feature set, explaining about 35% of the variance in transaction timing. To improve the model's performance, several approaches could be considered:

- Feature Engineering (add day of week, seasonal patterns, other interactions)
- Model Complexity (Explore non-linear relationships using polynomial features, consider time-series)
- Data Enhancement (include other data like weather, geographic factors)

This analysis suggests that while our current model provides valuable insights for marketing timing, there's significant potential for improvement through additional feature engineering and more sophisticated modeling approaches.

c. Business Impact and Benefits

The implemented model, despite its moderate predictive power ($R^2 = 0.354$), provides significant business value by transforming random promotional timing into a data-driven strategy. The model's predictions, while having an average error of 4.12 hours (MAE), represent a structured approach to marketing timing that was previously non-existent. As illustrated in the visualisation below, the model generates specific, actionable recommendations for different customer segments.

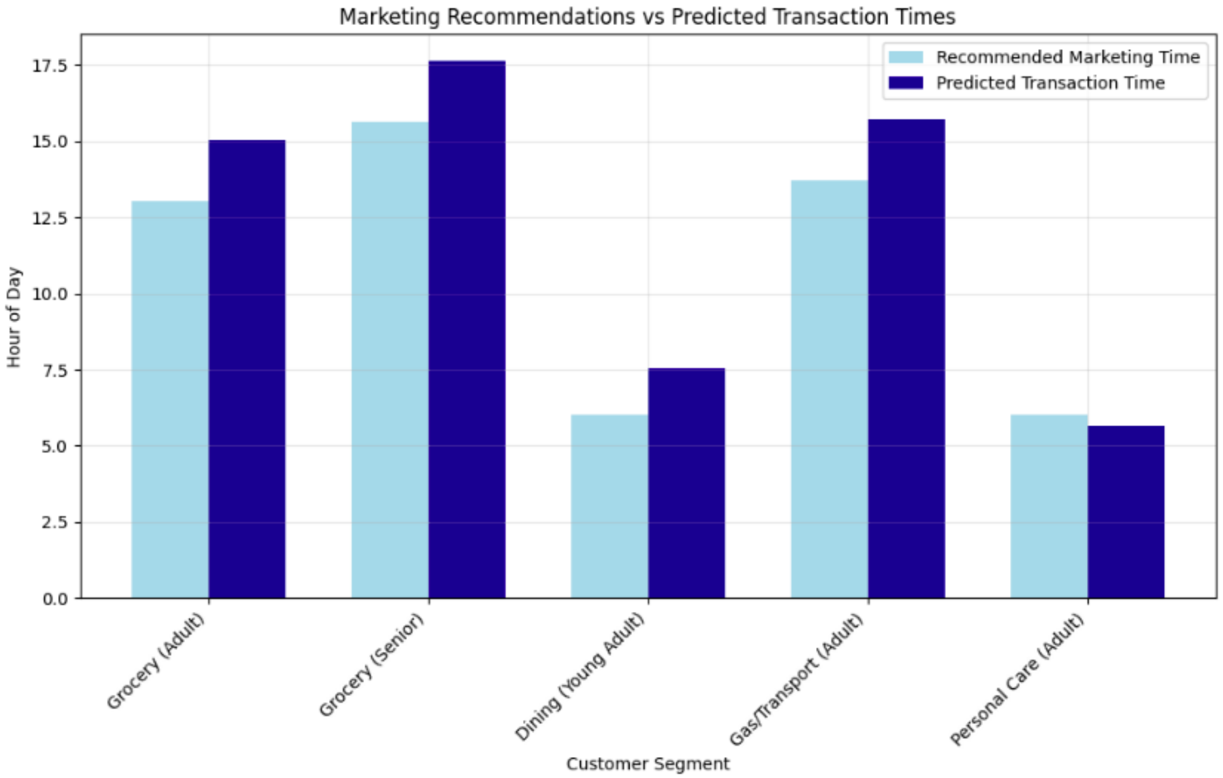


Figure 7: Timing recommendations for different segments

This initial version serves as a proof of concept, demonstrating that transaction timing patterns exist and can be leveraged for marketing purposes. Future iterations can incorporate additional data sources and more sophisticated modeling techniques, but the current model already provides actionable insights for improving marketing effectiveness.

d. Data Privacy and Ethical Concerns

The implementation of transaction-based marketing optimization requires careful handling of sensitive financial data. The dataset encompasses a wide range of personal and financial information, including credit card and account numbers, transaction amounts and timestamps, merchant locations, and customer demographic information. This wealth of sensitive data necessitates robust security measures throughout the entire data pipeline. Any real-world deployment would require comprehensive security protocols, including end-to-end data encryption, secure storage and transmission protocols, regular security audits, strict access controls, and thorough data retention policies.

The analysis revealed several significant ethical considerations that require careful attention. The most prominent concern is the substantial demographic disparity in our dataset, with 96.5% of transactions coming from urban areas compared to only 3.5% from rural areas. This imbalance raises important questions about representation and potential model bias. The risk of developing models that systematically favor certain demographic groups over others must be carefully managed to ensure fair and equitable treatment of all customers.

To address these concerns, I recommend implementing a comprehensive set of mitigation strategies. Data protection measures should include robust anonymization techniques and regular privacy impact assessments to ensure compliance with relevant regulations such as GDPR and CCPA. The implementation of fairness metrics in model evaluation and continuous monitoring of performance across different demographic groups will help identify and address any emerging biases. Building and maintaining customer trust requires transparent communication about data usage, clear consent mechanisms, and established channels for customer feedback.

Looking forward, the evolution of this system must consider several key factors. Privacy regulations and compliance requirements continue to evolve, as do customer expectations regarding data privacy. Advances in privacy-preserving machine learning techniques may offer new opportunities to balance marketing effectiveness with privacy protection. Regular reviews and updates to privacy and security protocols will be essential to maintain alignment with both technological capabilities and ethical standards.

The ethical implications of personalized marketing timing extend beyond simple privacy concerns. The ability to predict and influence customer behavior through targeted timing of promotions raises questions about autonomy and manipulation. Therefore, it's crucial to establish clear boundaries and guidelines for how this technology will be used, ensuring that customer benefit remains at the forefront of implementation decisions. By maintaining strong ethical standards and robust privacy protections while pursuing marketing optimization, we can build a system that serves both business interests and customer needs responsibly.

7. Conclusion

The development and implementation of a transaction timing prediction model represents a significant step forward in optimizing credit card reward promotions. Through careful analysis of over 816,000 transactions from 886 unique accounts, we have successfully created a model that can predict transaction timing with an average error of 4.12 hours, demonstrating meaningful improvement over random promotional timing. The Linear Regression model, selected for its balance of performance and simplicity, explains approximately 35.4% of the variance in transaction timing patterns, providing actionable insights for marketing strategy optimization.

The analysis revealed distinct temporal patterns across different customer segments and transaction categories, with notable variations between weekday and weekend behaviors, as well as between different times of day. These patterns, while not capturing all transaction timing variability, offer a structured foundation for moving away from random promotional timing toward a more data-driven approach. The model's ability to differentiate optimal marketing times for different customer segments and transaction categories provides valuable guidance for targeted marketing campaigns, potentially improving customer engagement and promotion effectiveness.

However, this project also highlighted several important areas for future development. The significant urban-rural disparity in the dataset (96.5% urban versus 3.5% rural transactions) suggests opportunities for improving data collection and model fairness. Future iterations could incorporate additional features such as seasonal patterns, weather data, and geographic factors, potentially improving the model's predictive power. Furthermore, the integration of more sophisticated machine learning techniques could help capture complex patterns that our current linear approach may not fully address.

The implementation of this model must carefully balance marketing effectiveness with privacy considerations and ethical implications. While the current model demonstrates the feasibility and value of transaction timing prediction, its deployment should be accompanied by robust privacy protection measures, regular bias monitoring, and clear customer communication protocols. By maintaining this balance, the bank can leverage data-driven insights to enhance customer experience while building trust and ensuring fair treatment across all customer segments. This initial implementation serves not only as a practical tool for optimizing promotional timing but also as a foundation for future developments in personalized banking services.