Executive Summary

Multiple Linear Regression Analysis

of Citi Bike Ridership Data

August 28, 2023

Prepared by: Javier Lopez, Data Analyst

For: Vice President of Data Science

Introduction

Problem Statement

The problem addressed in this study revolves around the need to predict Citi Bike ridership demand accurately using a comprehensive data-driven approach. Citi Bike, a bike-sharing system operating in New York City, faces the challenge of efficiently allocating resources, managing ridership, and optimizing operations to meet fluctuating demand. The dynamic nature of ridership, influenced by various factors like weather, seasonal patterns, and external events, necessitates a predictive model that captures these complexities and provides actionable insights for decision-making. The central research question that drives this endeavor is: "Can Citi Bike ridership demand be predicted using a multiple linear regression model based on market research data?"

Hypothesis

Given the intricate interplay of factors affecting Citi Bike ridership demand, the hypothesis of this study is that by employing a combination of advanced analytical techniques, including Ridge regression and time series modeling with ARIMA, it is possible to develop a predictive model that accurately forecasts ridership demand. Specifically, it is hypothesized that Citi Bike ridership demand cannot be accurately predicted through a multiple linear regression model based on market research data, with an accuracy surpassing 70%. On the contrary, the alternative hypothesis contends that such a model can indeed achieve a prediction accuracy exceeding 70%. The hypothesis asserts that this comprehensive approach will lead to a predictive model that not only achieves high accuracy in forecasting Citi Bike ridership demand but also offers valuable insights into the underlying dynamics of demand variations.

Methodology

The data-analysis process aimed to predict Citi Bike ridership demand through a systematic and multifaceted approach. The analysis began with a comprehensive exploration of the dataset using various techniques to uncover patterns, dependencies, and influential factors impacting ridership. The key steps of the data-analysis process are summarized below:

Data Collection and Preprocessing:

The analysis utilized historical ridership data from Citi Bike, encompassing attributes such as date, weather conditions, and pandemic periods. Data preprocessing involved handling missing values and converting relevant variables like date and counts to appropriate formats for analysis.

Seasonal Decomposition:

Seasonal decomposition was employed to understand underlying patterns within the time series data. This technique allowed the separation of the series into trend, seasonal, and residual components, providing insights into recurring temporal dynamics.

Log-Transformation:

To achieve a balanced residual distribution and improve modeling accuracy, log-transformation was applied to the residuals. This transformation mitigated the impact of extreme values and enhanced the model's robustness.

Autocorrelation and Partial Autocorrelation Analysis:

Autocorrelation and partial autocorrelation plots were used to identify temporal dependencies in the time series. These plots helped reveal lagged correlations, aiding in understanding repeating patterns and informing modeling decisions.

Dickey-Fuller Test for Stationarity:

The Dickey-Fuller test was conducted to assess the stationarity of the time series data. This test is crucial for time series analysis, as it determines if the series is stationary or not, a fundamental aspect of accurate modeling.

Ridge Regression Modeling:

Ridge regression, a form of linear regression with L2 regularization, was employed to build a predictive model for Citi Bike ridership demand. This technique addressed multicollinearity and overfitting by introducing a regularization term to the cost function.

Grid Search for Optimal Regularization Parameter:

A grid search was conducted to determine the optimal value for the regularization parameter (alpha) in the Ridge regression model. This process aimed to strike a balance between model complexity and generalization.

Evaluation Metrics:

Various evaluation metrics, including Accuracy Rate, R-squared, Mean Absolute Percentage Error (MAPE), and Root Mean Square Error (RMSE), were calculated to assess the performance of the Ridge regression model and its accuracy in predicting ridership demand.

Forecast Horizon Selection:

An iterative search approach was used to identify the best forecast horizon that minimized prediction errors. Different forecast horizons were evaluated to determine their impact on the model's accuracy.

Visualization of Actual vs. Predicted Values:

The analysis included graphical representations that showcased the alignment between actual and predicted ridership demand values. Confidence intervals were incorporated to provide a measure of prediction uncertainty.

Comparison of Results:

The results obtained from the Ridge regression model were compared to those of ARIMA time series models in terms of accuracy, MAPE, and RMSE. This comparison provided insights into the effectiveness of each modeling approach.

Feature Selection and Model Refinement:

A step-back methodology was used for feature selection, refining the model by including only relevant predictors. The impact of this refined model on accuracy and RMSE was assessed.

Impact of Forecast Horizon:

Different forecast horizons were tested to determine their influence on the model's predictive performance. The R-squared, MAPE, RMSE, and Accuracy Rate were evaluated to quantify the effect of different horizons.

Implications and Recommendations:

The analysis provided insights into the factors influencing Citi Bike ridership demand and recommended strategies for resource allocation, marketing, and infrastructure improvements. Future directions for study were also outlined, including advanced feature selection and exploring dynamic time series models.

In summary, the data-analysis process encompassed a wide range of techniques, from exploratory data analysis and transformation to advanced modeling and evaluation. The combination of Ridge regression and time series modeling allowed for a comprehensive understanding of the dataset's complexities and the development of accurate predictive models for Citi Bike ridership demand.

Key Findings

The analysis investigated ridership patterns in an urban bike-sharing system, utilizing various methodologies including data exploration, time-series analysis, and predictive modeling. The objective was to understand the factors influencing ridership and create accurate demand prediction models. This summary provides a concise overview of the observed trends, important predictors, and model efficacy.

Data Exploration

- **Weekly, Monthly, and Yearly Trends**: Rolling mean calculations revealed an accumulative trend across successive years with a temporary downturn during the initial pandemic period.
- **Seasonal Distribution**: Summer held the majority share at 31% of total trips, while Winter contributed 16%.
- **Correlation Analysis**: Temperature and hour had the most substantial positive correlation with ridership, while humidity had the highest negative correlation.
- **Holiday Influence**: Holidays triggered a decline in Manhattan's ridership, an uptick in Brooklyn, and relatively unaffected patterns in Queens.
- **Rider Type Distribution**: Despite the overall decrease in Manhattan's ridership, casual riders increased across all boroughs during the same period.
- **Covid-19 Impact**: Significant shifts in ridership occurred during morning and evening rush hours. Morning rush hour trips declined during lockdown, and evening rush hour trips increased.
- **Weekday and Pandemic Period**: Ridership distribution shifted during lockdown and reopening phases, with variations from pre-pandemic patterns.

Time-Series Analysis

- **Seasonal Decomposition**: Revealed an annual cyclic trend in the data with a lack of discernible patterns in the trend component.
- **Autocorrelation and Partial Autocorrelation Plots**: No significant lags in autocorrelation, but strong partial autocorrelation at lag zero suggested auto-regressive component.
- **Dickey-Fuller Test**: The dataset showed stationary behavior, confirming consistent patterns over time.
- **Variance Inflation Factor (VIF)**: Multicollinearity concerns were addressed through feature selection, improving model robustness.

Model Development

ARIMA

- **Initial ARIMA Model**: Achieved an Accuracy Rate of 73.3%, MAPE of 26.69%, and RMSE of 905.44.
- **Optimal Forecast Horizon**: Six months forecast horizon improved Accuracy Rate to 81.66% and reduced AIC by 15.08%.
- **ARIMA Forecast with 6-Month Horizon**: Actual versus predicted values aligned well, capturing data patterns with confidence intervals.

Multiple Linear Regression

- **Ridge Regression**: Employed Ridge regression to address multicollinearity and overfitting.
- **Initial Ridge Model**: Achieved an Accuracy Rate of 73.23%, R-squared of 82.60%, MAPE of 26.77%, and RMSE of 893.52.

- **Feature Importance**: 'Bike_counts_log' and 'pandemic_period' were influential predictors in the model.
- **Second Ridge Model**: Accuracy Rate of 73.28% and RMSE of 893.52, relying on twelve predictors.
- **Optimal Forecast Horizon**: 6-month horizon improved R-squared to 93.97%, MAPE to 17.66%, and RMSE to 494.17.
- **Residual Analysis**: Residuals exhibited deviation in extreme observations, indicating room for improvement.

These findings provide insights into ridership patterns, influential variables, predictive models' performance, and the impact of forecast horizons. The combination of data exploration, time-series analysis, and model development offers a comprehensive understanding of Citi Bike ridership trends and predictive accuracy.

Limitations

The analysis of Citi Bike ridership data incorporated a range of techniques and tools to uncover insights and optimize predictive accuracy. However, it's essential to recognize that these techniques and tools come with their own set of limitations, which should be acknowledged to provide a comprehensive understanding of the analysis process.

Log-Transformation and Residuals:

While log-transformation is valuable for normalizing residuals and addressing variance issues, it might not always achieve perfectly normalized distributions. This can result in residual patterns that are not centered around zero, potentially affecting the accuracy of model

assumptions. Additionally, the interpretation of coefficients becomes more intricate due to the transformed scale, posing challenges in conveying results effectively.

Autocorrelation and Partial Autocorrelation Plots:

Autocorrelation and partial autocorrelation plots are useful for identifying temporal dependencies, but they might not capture complex interactions involving multiple lags. In cases with intricate lag relationships, these plots could overlook significant patterns, potentially leading to incomplete modeling decisions.

Dickey-Fuller Test:

While the Dickey-Fuller test is vital for assessing stationarity, it has limitations in pinpointing the specific causes of non-stationarity or the presence of structural breaks. It can determine if a series is stationary, but it might not offer insights into the underlying reasons for temporal dependencies, limiting the depth of understanding.

Variance Inflation Factor (VIF) for Feature Selection:

Utilizing VIF for addressing multicollinearity is effective, but it primarily focuses on linear relationships and might not fully capture complex interdependencies beyond linear correlations. This might lead to a selection of predictors that don't fully encapsulate the complexity of relationships among variables.

Forecast Horizon Optimization:

While selecting the best forecast horizon through iterative search enhances predictive accuracy, this approach might not account for abrupt shifts or changes in trends that could impact longer-term predictions. External events or unforeseen changes could render the chosen horizon less suitable for capturing evolving dynamics.

Recognizing these limitations provides a context for interpreting the results and making informed decisions. Addressing these limitations could involve exploring complementary techniques, employing more advanced models, or considering additional data sources to enhance the accuracy and robustness of the analysis outcomes.

Proposed Actions

Based on the results of the analysis, I recommend leveraging the insights gained from the Ridge regression model to inform decision-making and strategies related to Citi Bike operations and ridership management. The model's ability to highlight significant predictor variables and their relationships with ridership demand can aid in optimizing resource allocation, marketing efforts, and infrastructure improvements. The following actions are proposed to address the limitations listed above:

Data Preprocessing Enhancement:

- Explore alternative normalization techniques in addition to log-transformation to mitigate skewed distributions and variance.
- Investigate methods that address residual normalization more effectively, considering the limitations of log-transformations.
- Utilize complementary approaches alongside autocorrelation and partial autocorrelation plots to capture intricate temporal dependencies beyond lag analysis.

Non-Stationarity Understanding:

- Combine the Dickey-Fuller test with supplementary diagnostics to uncover specific causes of non-stationarity and potential structural breaks in the time series data.

- Investigate external factors or events that might contribute to non-stationarity, enhancing the interpretation of temporal dependencies.

Advanced Feature Selection:

- While Variance Inflation Factor (VIF) addresses multicollinearity, explore techniques like LASSO regression or Elastic Net to capture non-linear relationships among predictor variables.
- Implement more advanced feature selection methods that consider both linear and non-linear dependencies, leading to a more comprehensive predictor subset.

Adaptive Forecast Horizon:

- Combine iterative search for forecast horizon optimization with dynamic modeling or scenario analysis to account for sudden shifts and evolving trends in ridership demand.
- Consider incorporating external data sources or events that might influence ridership patterns beyond the forecast horizon.

These proposed actions aim to refine the analytical process by addressing specific limitations and enhancing the robustness of predictive modeling. By proposing these actions, the study strives to provide stakeholders with more accurate insights, enabling them to make informed decisions and strategies for managing Citi Bike ridership effectively.

Expected Benefits

The study on Citi Bike ridership analysis and predictive modeling offers several anticipated benefits that can significantly impact decision-making and operational strategies. While the outcomes are contingent on the accuracy of the model and the insights drawn, the study envisions specific and quantitative benefits across various domains. By leveraging

advanced predictive modeling techniques like Ridge regression and time series analysis, the study aims to enhance the accuracy of Citi Bike ridership predictions. An expected improvement of at least 5% in prediction accuracy is anticipated compared to baseline methods. Accurate ridership predictions facilitate optimized resource allocation for Citi Bike operations. With enhanced accuracy, the study aims to reduce underutilization of bikes and stations by at least 10%, resulting in more efficient distribution and management of resources.

The study also expects improved revenue projection for Citi Bike services. Anticipated is a reduction of prediction errors in revenue projection by at least 8%, leading to more reliable financial planning. Additionally, insights gained from the analysis have the potential to transform marketing strategies. The study aims to identify at least two rider segments with distinct behaviors, thereby enabling targeted marketing campaigns and potentially increasing user engagement by 15%. Effective operational planning is another benefit, as the ability to predict ridership trends aids in proactive measures. The study's insights could assist Citi Bike in reducing station downtime for maintenance by at least 10%, minimizing disruptions for riders. The efficient allocation of resources and proactive maintenance directly translate into cost savings, with an expected reduction of operational costs by 12%.

Strategic infrastructure development can also be informed by accurate ridership predictions. The study aims to increase the accuracy of station expansion predictions by at least 20%, ensuring that new stations are strategically placed where demand is projected to be high. Moreover, the insights derived from the analysis will empower stakeholders to make informed decisions. An increase of at least 15% in decision-makers' confidence levels is expected when making strategic choices based on analysis outcomes. The study's accurate predictions and insights can provide Citi Bike with a competitive edge in the bike-sharing market. An expected

increase in Citi Bike's market share by at least 5% is anticipated due to more reliable service and user-friendly features informed by analysis outcomes. Better operational planning and service availability translate into an improved user experience, with an expected increase of at least 8% in user satisfaction ratings due to reduced service disruptions and enhanced availability.

In summary, these expected benefits underscore the potential impact of the study on Citi Bike's operational efficiency, financial planning, and overall user experience. However, it's important to note that the extent of these benefits will be contingent on the accuracy of the predictive model and the insights drawn from the analysis.

Sources

No external sources were used in the drafting of this executive summary.