**UC DAVIS MSBA**

**BAX 452 Final Project Report**

**Winter 2022**

**Anna Gaudette**

**Priyanka Priyanka**

**Akansha Bansal**

Contents

[1.1 Executive Summary 3](#_Toc98796859)

[1.2 Business Background and Potential Impact 4](#_Toc98796860)

[1.3 Data and Approach 4](#_Toc98796861)

[1.4 Analysis 5](#_Toc98796862)

[1.5 Recommendations and Business Value 10](#_Toc98796863)

[1.6 Summary and Conclusions 11](#_Toc98796864)

[1.7 References 11](#_Toc98796865)

# 1.1 Executive Summary

The goal of this exercise is to examine bike rentals within a specific location to determine whether forecasting at a granular level will isolate factors influencing two customer types: casual users (first-time users) and registered users (subscribed, long-term customers).

A variety of machine learning techniques (described in detail in section 1.4) were applied to collected data to isolate factors useful for predicting rentals across months for financial and operational forecasting and by times of day for more granular inventory and maintenance planning. The initial approach examined total user forecasting based on a variety of variables. Through the resulting analysis, we found that the two distinct user groups, Registered and Casual, should be examined separately due to different profiles. End results findings

Based on our findings, we believe that we can create a financial and operational forecasting tool that accounts for differences in predictive factors between these two user groups. Other uses beyond financial modeling include inventory and maintenance planning by isolating factors that indicate lower short-term and long-term use, reducing the amount of down time during peak periods. Additionally, we believe that we can utilize different marketing timing between the two groups to target them when they are mostly likely to engage with the product.

# 1.2 Business Background and Potential Impact

The bike rental market serves a variety of customers – recreational, health-conscious, or those pursuing non-fossil fuel sources of transportation. The cost to operate is mostly upfront asset purchases with relatively high profit margins. Reports indicate that many local governments are partnering with other companies within the industry to offer subsidies. Depending on the customer’s use, bike rentals could be as a form of recreation or as a form of transportation used as an alternative to public transit, personal automobiles, or other. As a result, competition runs the gamut from other bike rental companies to public transportation to on-demand ride hailing services.

Forecasting methods are usually applied at a high level by examining levels, trends, and seasonality for inclusion into financial and operational forecasts. Our use case is to develop a forecasting tool based on collected data to predict total bike rentals, casual (first time) users, and registered (or subscribed users) for incorporation into financial forecasting.

# 1.3 Data and Approach

Data was sourced from Fanaee and Gama, “Event labeling combining ensemble detectors and background knowledge”, which combined publicly available bike sharing system data from Capital Bikeshare’s (<http://capitalbikeshare.com/system-data>) system and hourly weather information from i-weather.com.

The dataset consisted of 17,379 hourly observations of 31 total regressors (continuous, discrete, and categorical) covering weather and other situational variables. The variables utilized in varying Machine Learning approaches (discussed below) and their associated description are as follows:

* Month - Categorical
* Hour – Cyclical Variables (1)
* Holiday – Categorical
* Weekday – Categorical (Sunday – Saturday)
* Workingday – Categorical (not a holiday or weekend)
* Weathersit – Categorical (1 - Clear, 2 - Cloudy, 3 - LightPrecip, 4 - HeavyPrecip)

- 1: Clear, Few clouds, Partly cloudy, Partly cloudy

- 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist

- 3: Light Snow, Light Rain /Thunderstorm/Scattered clouds, Light Rain/Scattered clouds

- 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog – (“HeavyPrecip”)

* Temp –normalized temperature
* Atemp – feeling temperature, Celsius
* Hum – normalized humidity
* Windspeed - windspeed
* Cnt – count of total rental bikes
* Casual – count of first time, non-registered users
* Registered – count of subscribed users (monthly pass, etc.)

After initial exploratory data analysis, several variables were eliminated, including date, year, and season. Seasons were deemed redundant to the monthly variables in addition to having too high a level that obscured transition periods between seasons.

# 1.4 Analysis

Multiple techniques were used to examine the dataset to determine factor significance, including OLS Regression, Lasso, Lasso with Cross Validation, and Random Forest. These techniques were applied to data for both registered users and casual users in order to analyze whether certain factors were better predictors for different customer types.

OLS Regression:

The first analysis applied Linear Regression to the reduced (as described above) variable set. Given the nature of the data and the inclusion of a significant number of variables, the adjusted R-squared for the model was 1 and many variables has inflated t-values, suggesting the presence of multicollinearity. For Registered users, the following variables were not statistically significant – Mondays, Wednesday, ambient temperature, humidity, temperature, windspeed, and hour. For Casual users, these variables were not – Mondays, temperature, windspeed, and hour. Given the correlation between month/hour of the day and various weather factors, the majority of these are unsurprising.

Lasso Regression:

To help mitigate the presence of multicollinearity indicated in the Linear Regression, we performed Lasso Regression techniques on training data (70% of data randomly selected) with the goal of removing all but one of a set of highly correlated variables. The out of sample R-squared was then calculated as a comparison point to the in-sample training data.

For the Registered Users, 12 regressors were selected by the lasso regression technique. Of these, temperature (Graph 1), clear weather, the October and September dummy variables, and whether the day was a working day were positively correlated with registered users. The coefficients for hour, Monday, light precipitation, and the January through March months were all negative, suggesting a relatively low user base for early and late hours (hours was cyclically transformed), slight rain conditions, the first day of the work week, and several of the winter months. Overall R-squared was relatively low at 35.14% within sample and 35.53% on the reserved test data.

Chart, histogram, scatter chart

Description automatically generated

*Graph 1*

For the Casual Users, Lasso Regression isolated only 6 variables of the 31 regressors with a positive impact for only temperature and clear conditions, suggesting that casual users tend to engage the service for warm days with no rain. Hours, whether the observation was a working day, and light precipitation were all negative coefficients. Extrapolating from these and comparing to the Registered users, casual users had a negative impact from the observation being a working day (Graph 2), but had similarly negative coefficients to Registered users for the hour, whether the weather was clear, and whether there was light precipitation although the magnitudes varied (Graph 3). Overall R-squared was better at 52.09% within sample and 52.12% on the reserved test data.

Chart, box and whisker chart

Description automatically generated

*Graph 2*

A picture containing text, device, screenshot

Description automatically generated

*Graph 3*

Lasso Regression with Cross Validation:

Despite the relatively close in sample and out of sample R-squared values, we performed Lasso Regression with Cross Validation to examine variable stability in terms of inclusion in the two models. For registered users, R-squared showed negligible improvement at 35.91% within sample and 35.86% on the reserved test data. The number of coefficients, however, increased to 28 of the 31 regressors. Only cloudy weather, heavy precipitation weather, and Thursday were removed by the Lasso with CV. A similar pattern occurred with the Casual users - the number of coefficients increased to 28 (excluding cloudy weather, heavy precipitation weather, and Monday) with a marginal improvement in R-squared (56.98% in sample, 57.11% out of sample). The results indicate a lack of stability in the inclusion of the coefficients between folds.

Random Forest:

Bike rentals are likely a function of multiple decisions points, for example deciding to ride a bike in the morning instead of the night, followed by a decision point around weather conditions. Given this, the final technique we used to explore the data was Random Forest and we explored built-in interactions between variables. We then isolated variable importance and examined the differences in the first three variables for the two user types.

For the Registered users, variable importance ranked hours the highest, followed by whether the day was a working day, and then temperature. In contrast, Casual users showed that ambient temperature was the higher ranked importance factor (Graph 4) followed by working day, and then hour. In contrast to the Lasso Regression, there was a larger difference in accuracy between the in sample and out of sample data.

Chart, histogram, scatter chart

Description automatically generated

*Graph 4*

After examining the four techniques outlined above, we selected Lasso regression as the final approach due to its relative simplicity and parsimony in the number of selected variables.

# 1.5 Recommendations and Business Value

Based on the above analysis, we recommend that financial and operational forecasting should be performed separately on the two user groups rather than at the consolidated user base level. In addition to using this analysis to understand key drivers of use and the differences between the two different types of consumers, revenue profiles are different between the two groups, creating a mixed revenue forecast. While the two groups both indicate a positive correlation in use with higher temperatures and clear weather, Registered users have a positive correlation with working days while Casual users have a negative correlation with working days. Further analysis needs to be performed to determine whether Casual users are primarily within the recreational use category and Registered users are more in the daily transportation category. Understanding these behavioral patterns will help in targeting certain customer segments depending on associated profit margins.

Following from this analysis, several additional concepts can be explored, including providing the ability to create more accurate inventory planning. This analysis provides information around when use cycles are likely to be lower and bikes can be taken out of rotation for short term general maintenance (heavy precipitation days, low temperature days, periodic intervals throughout the day), when extra inventory might be required ahead of certain events, and when asset replacement cycles should occur (i.e., directly prior to the seasonal uptick in use after March). In addition to revenue planning, knowing when asset replacement should occur provides more accuracy for cash flow timing. Projecting optimal capacity will help balance between not missing out on sales and incremental holding costs associated with maintaining too much of an unneeded fleet at certain times of year.

Identification of different factors as drivers of differences between the two types of consumers may allow for targeted ads at specific intervals to acquire new Casual customers to reduce advertising costs at lower use times or to test discount rates for casual customer acquisition at periods where they are most likely to adopt the product (weekends, holidays, warm temperature days).

Lastly, the more granular customer type level can help provide a use case for working with local city personnel to obtain subsidies for reducing car traffic during peak transport periods.

# 1.6 Summary and Conclusions

In summary, this analysis indicates that we should move to a two-customer model for future forecasting and planning. We recommend incorporating this analysis into financial projections, inventory and maintenance analysis, and marketing efforts. Additionally, pursuing even more granularity by collecting geo-location specific tracking for local vs city-wide planning initiatives may provide more insight into customer patterns and better demand prediction to ensure minimal lost rental opportunities.

# 1.7 References

1. Cox. (2006). Speaking Stata: In Praise of Trigonometric Predictors. The Stata Journal, 6(4), 561–579. <https://doi.org/10.1177/1536867X0600600408>
2. Fanaee-T, & Gama, J. (2013). Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, *2*(2-3), 113–127. <https://doi.org/10.1007/s13748-013-0040-3>.
3. https://www.ibisworld.com/united-states/market-research-reports/bike-rental-industry/