McCombs School of Business

METROBIKE AUSTIN TRIPS

Intro to Machine Learning

Group Project

Chyavan Mysore Chandrashekar	CM65624
Pratik Gawli	PBG397
Rukh Agha	MSA3453
Spoorthi Anupuru	SA56643
Tanushree Devi Balaji	TB33857

1. Data Collection and Processing

As a team, we started by individually exploring the various data sources on the internet. Each of us proposed data sets and data sources that were interesting and amusing in terms of answering a business question. We scouted and collated sources that provided data in various forms - public data sets like https://data.gov/, APIs like https://www.googleapis.com/youtube/v3, and web scrapping libraries like https://github.com/davidteather/TikTok-Api. Some of the interesting business questions we isolated at the end of our first meeting were,

1. What should be the ideal platform and content that a Marvel character should be launched on to maximize engagement and commercial value?
2. Does a particular player's transfer cause a positive or negative impact on the soccer association's rating, and in turn funding and sponsorship?
3. How likely is a movie to be nominated or win an Oscar?
4. **How can we estimate the supply standards of the Austin MetroBike stations and identify the locations and seasons of low traffic for ideal promotions?**

We finalized the last business question and collected the data from the government website - https://data.austintexas.gov/Transportation-and-Mobility/Austin-MetroBike-Trips/tyfh-5r8s. The data consisted of 1.69 million rows where each entry corresponded to a MetroBike trip of a customer. We removed the incorrect and null entries. The data consisted of 13 columns out of which we considered columns **Trip Duration Minutes**, **Checkout Kiosk**, **Checkout Date**, and **Membership Type**, and aggregated the rest of the columns to get a new column that corresponded to **Number of Trips**.

We then transformed the columns as follows,

1. Checkout Kiosk – We mapped the location of each Kiosk into 2 new columns corresponding to the "Latitude" and "Longitude" of the Kiosk. Some of the locations could not be transformed into latitude and longitude by the `geopy.geocoders` library and required manual updation.
2. Membership Type – 77 total memberships were found among the 1.6 million entries where most except the top 9 entries were outliers having only 1 or at most a few rides. We removed these entries and considered only the top 9 entries.
3. Temperature – We added another column called Temperature which contained the average monthly temperature in Austin for each entry
4. Trip Duration Minutes – We converted the trip duration into a binary qualitative variable where it could have a value of 1 signifying a "Long trip" or 0 signifying a "Short trip"
5. Months – We converted the months from a numerical column to dummy variables.

After the transformation, the data consisted of 34k rows which were further divided into training and test sets randomly to test the fit with the Validation Set approach.

2. Exploratory Analysis

We performed exploratory analysis by calculating the correlation matrix and looking for predictor trends for various predictors. The correlation matrix did not provide us with any information as the data was convoluted. The plots for certain predictor trends showed the expected relationships between the response (# of Trips) with predictors latitude, longitude, and if the trip was short or long.

3. Analysis

We performed multiple analyses on the data using both parametric and non-parametric models. The table below summarizes the approaches and their accuracy in terms of Test RMSEs.

Sl. No.	Model	RMSE
1.	Multiple Linear Regression	402.4
2.	Forward Regression	382.3
3.	Log Regression	252.7
4.	K Nearest Neighbors	48.50
5.	Boosting	32.10
6.	Random Forests	41.06
7.	**Bagging**	**29.45**

Out of all the models, Bagging gave us the optimal test-RMSE of 29.45, i.e. the predictions with our best fit model had a standard deviation of 29.45 trips from a station. For a range

4. Insights and Conclusion

Linear models did not perform well due to the lack of clear-cut linear relationships. However, non-parametric methods performed much better given the flexibility they bring in terms of shaping the predictive curve.

Of all the models, bagging yielded the best results and it makes sense to utilize a non-parametric bagging model built based on location, month, and membership type to predict prospective trip volumes. This information can be utilized in two main ways:

- Supply Management - Plan in advance for high traffic months (eg: March or October) for specific customer categories at high-performance locations
- Marketing - Identify low-performance months/locations with a good customer pool and see if sales can be boosted through advertisement campaigns

Looking Back: Another enhancement for the non-parametric models would be to use the most recent 12 months as a hold-out set for validation, rather than a random split to more effectively attempting a time series analysis (which was out of the scope of our lectures). An alternative would be to run non-parametric models without the ‘year’ predictor to “anonymize” months as a standalone feature and keep them more categorical. This is included in the appendix for bagging.