Philadelphia’s Indego Bike Share Service 2018 Trip and Station Data Analysis

Leonard Armstrong Jackole Meyers  
IST-707 Data Analytics  
Syracuse University

# Introduction

The first model for sharing a common group of bicycles across a community started in Amsterdam in 1965 by a group called Provo. The Provo program was a short-lived and failed experiment that saw many of the bicycles either stolen or abandoned in canals.[[1]](#footnote-1) Perhaps the Provo program was a half-century before its time. In the past decade technological developments have made it possible for many communities to create secure, environmentally-friendly bike sharing services, disrupting established public transportation models.[[2]](#footnote-2)

Bike share popularity is growing. There are currently bike share services serving 56 major US cities. The National Association of City Transportation Officials reports that, when compared to 2016, 2017 saw a 25% increase in bike share trips and a 135% increase in the number of provided bikes. 2017 also saw an 84% increase in ridership for the Indego service in Philadelphia.[[3]](#footnote-3)

Indego is the public bike share service owned by the City of Philadelphia with naming rights sponsorship by Independence Blue Cross, hence the “*Inde*” in Indego. Each Indego station consists of a high-tech kiosk and a set of bike docks operating 24 hours a day, seven days a week. Each kiosk manages anywhere from 11 to 57 docks. In 2018 Indego served almost two-thirds of a million trips from over 1400 bikes available at 130 kiosks.

A typical Indego station with kiosk, docks and bikes is shown in Figure 1.

Figure 1- An Indego Kiosk Station and Bikes

Indego has a multi-tier price model with different pass types to meet a variety of traveler needs. Available Indego pass types and their associated costs are detailed in **Figure 2**.

| Plan | Term | One-time Cost[[4]](#footnote-4) | Per-Trip Costs |
| --- | --- | --- | --- |
| Indego30 | 1 month | $17 | Unlimited one-hour trips at no additional cost.  $4/hour after the first hour of a trip. |
| Indego365 | 1 year | $156 | Unlimited one-hour trips at no additional cost.  $4/hour after the first hour of a trip. |
| Day Pass | 1 day | $10 | Unlimited 30-minute trips at no additional cost.  $4/30-minutes after the first 30 minutes of a trip. |
| IndegoFlex | 1 year | $10 | $4/hour for all trips. |
| Walk-up[[5]](#footnote-5) | 1 trip | - | $4/30-minutes |

Figure 2- Indego Price Model Summary

A general study of Indego 2018 trip and station data is performed, and its results are reported here. The study includes exploratory data analysis, geographic mapping, rule mining, clustering, and predictive classification models.

# About the Data

Since it is a public service, Indego makes its trip and station data free, open, available and documented at <https://www.rideindego.com/about/data/>. Subsets of Indego data may also be found on data clearinghouse sites such as data.world (<https://data.world>).

## About the Trip Data

Indego makes trip data available in CSV files where each available file contains three months, one calendar quarter, of trip data. Quarterly data is currently available since Indego’s launch in 2015 Q2 through the most recently completed quarter, 2018 Q4. The fields available in the CSV files have evolved as new services and capabilities have been added or altered. Over the four quarters of 2018 one new field, bike\_type, was added to the CSV structure due to the introduction of electronic bikes to the fleet. Indego 2018 trip data can be summarized as:

* Trip identification information
* Trip duration information
* Trip station information – both source and destination stations
* Bike information
* Pass type used

The CSV structure contains some redundant information. For example, there is a field called trip\_route\_category that contains values of “One Way” and “Round Trip”, but these values can be determined by comparing the start (source) station to the end (destination) station, also identified in the CSV.

This study used data from the entirety of 2018. Therefore, the complete trip data set was assembled from four separate CSV files and combined into one internal R data frame consisting of all trips, as illustrated in Figure 3.

A close up of a logo

Description automatically generated

Figure 3- Trip Data Read Model

## About the Station Data

Alternatively, Indego station data is made available in either GeoJSON or General Bikeshare Feed Specification (GBFS) formats.[[6]](#footnote-6) There is also a simplified CSV station table available, but that table lacks many of the fields available in the GeoJSON and GBFS formats. The station data in both the GeoJSON and GBFS data is made available live, in real time. A copy of the GeoJSON format was captured for use in this study on 18 February 2019 at 2:08PM, EDT. A summary of the available station data includes:

* Station identification information
* Station location information
* Kiosk structure information
* Kiosk status information
* Open dock availability
* Bike availability

While the JSON and GeoJSON formats allow for a deep hierarchy, the information provided by Indego’s GeoJSON instance was able to be transformed into a single flat data frame structure without any loss of information. Thus, both the trip data and the station data were able to be captured as R data frames. The station data read model is shown in Figure 4.

A close up of a sign

Description automatically generated

Figure 4- Station Data Read Model

## The Data Model

Figure 5 shows a model of the complete set of trip and station fields made available by Indego and their interconnectivity. These data entities have gray header blocks. Also included in the figure with blue headers are two desired data entities that were not made available by Indego: detailed bike information and rider information. Only bike\_id and bike\_type fields are included in the trip data. Additional bike information such as maintenance schedule, dates of release and retirement, and current location is not provided. Indego provides no rider information at all making ridership assessment and profitability very difficult information to ascertain. Even an anonymized rider ID would have been highly valuable.

In total and after data cleansing, the Indego data set provided information on 636,461 trips, 130 stations/kiosks, and 1,466 bikes. While a separate bike table was not provided, bike volume was computed from the set of unique bike identifiers available in the trip table.

A screenshot of a cell phone

Description automatically generated

Figure 5- Available and Desired Data Model

## Appending the Data Due to Unavailable Bike Types

Prior to 2018 Q3 all Indego bikes were of the same manual type. In 2018 Q3, Indego prepared for the addition of electric bikes to their fleet by adding a new bike\_type field to their trip data, requiring values of either standard or electric in this field. However, the Q1 and Q2 data files lacked this bike\_type field. Therefore, after reading the CSV files for 2018 Q1 and 2018 Q2, the data was updated to include a standard bike type for every record from those two files.

## Cleansing the Data

The process of data cleansing started before the data was provided to the public by Indego. Specifically, Indego lists the following data cleaning operations that were performed prior to making the data available.

* Staff servicing and test trips are removed.
* Trips under one minute in duration are removed.
* Trip lengths are capped at 24 hours.

Indego also notes “some short round trips or long trips may be the result of system or user error but have been kept in the dataset for completeness.”[[7]](#footnote-7)

Once all trip data was consolidated into a common data frame, a series of verifications were undertaken, and the data was updated as needed. First, a check was performed to ensure that every bike\_id was consistently paired with a single bike\_type (either standard or electric). All bikes were consistently typed so no updated was needed from this check.

Then, availability checks were performed to see what required data elements were non-existent. This check included searching the data for both NA and NULL values. There were 7947 records that contained NA values. All these values were located in the location fields, so these trip records were removed.

There was also inconsistency in the use of the single day pass model. Both "One Day Pass" and "Day Pass" were used in the passholder\_type field. No documentation existed that distinguished between these two values so all “One Day Pass” passholder\_type fields were changed to “Day Pass” making for a single, consistent value.

A review of station latitudes and longitudes in the trips data revealed one station whose latitude was incorrectly listed as a negative value when it should have been a positive value. The record was kept[[8]](#footnote-8), and its station latitude was updated back to a positive value.[[9]](#footnote-9)

After reading the station data from its GeoJSON source, the now-common NA checks were performed, showed no issues, and no station records were removed.

While the trips and stations data may have been internally consistent, the use of a 2-table relationship requires that consistency checks are made across the table’s relationship(s). Three consistency checks were made. First a check was made to ensure that all start stations had a corresponding station record in the stations table. Second a similar check was made to ensure that all end stations had a corresponding station record in the stations table. Finally, a check was made to assure that all stations in the station table were used for at least one trip in the trips table. 13,941 records were found in the trips table with start and/or end station values of 3023, 3036, 3122, and 3106.[[10]](#footnote-10) These four stations were not found in the stations table, so those records were removed from the trips table. All stations were determined to be in use for at least one trip.

## Exploratory Data Analysis

With the full and clean set of trip and station data on hand, a series of exploratory data analyses were made to listen to the story the data was telling.

### When Are People Riding?

The first set of exploratory analyses were undertaken to better understand the time patterns of Indego’s ridership at multiple levels of scope. Figure 6 is a histogram of the volume of bike trips taken each month in 2018. Each month is stacked with the pass type volumes used within the month. The trip volumes show a near-normal distribution peaking in July. There are two interesting things to note from the figure.

First, Indego30 is the most popular plan used for 2018 trips. Second, the Day Pass was introduced in April and effectively eliminated the Walk-up plan that was slated to be phased out. While Walk-ups were still used in very small percentages until the end of the year, this data shows that the Day Pass, available directly at the kiosks, provided a successful transition plan, satisfying the needs of most walk-up users.

A picture containing stationary, writing implement

Description automatically generated

Figure 6- Frequency of Trips by Month and Pass Type within Month

Next, Figure 7 shows 2018 Indego trip distribution in a histogram organized by day of the week with each day stacked by trip type volume. Again, we see a fairly normal distribution, this time peaking on Wednesdays.

This figure makes clear that one-way trip volumes far exceed round-trip volumes and that round trips, for the most part, have a more uniform distribution across all days of the week. A round trip is one where the start station and end station are identical whereas a one-way trip starts at one station and completes at another.

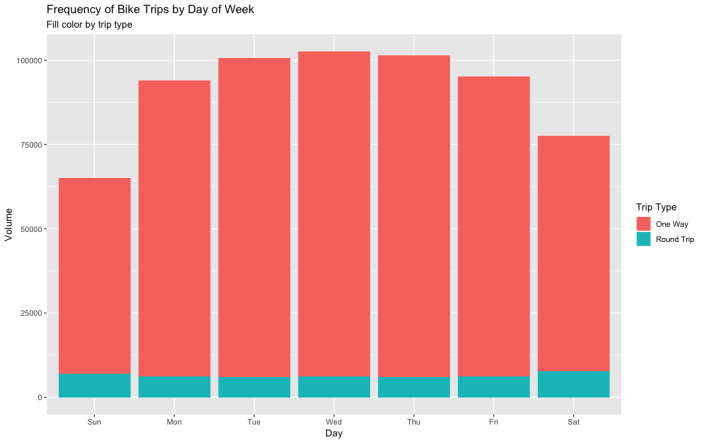


Figure 7- Frequency of Trips by Day of the Week and Trip Type Within Day

The imbalance of the trip type distribution and the pass type distribution lends to a concern that there may not be sufficient distinguishing factors for some of the machine learning algorithms.

Next, Figure 8 maps 2018 trips in a histogram organized by the hour of the day and stacked with day of the week coloring. Again, there are two interesting things to note from this histogram. First, at the macro level we do not see a normal distribution for this histogram. Instead we see a bi-modal distribution with peaks occurring in the 8AM hour and the 5PM hour, corresponding to trips to and from work for a normal business day. This implies that Indego may be heavily used for business travelers.

Second, the histogram contains an embedded normal distribution during the weekend. Stacked using the bottom two colors of the graph, Sunday is in yellow and Saturday is in a yellow-green color. Unlike the overall distribution, the weekend distribution peaks in the afternoon hours, between noon and 4PM.

A picture containing vector graphics

Description automatically generated

Figure 8- Frequency of Trips by Hour of the Day and Day of the Week Within Hour

### Station Usage

Next station distribution and usage is analyzed. An outline of Philadelphia’s Y-shaped city boundaries and the clustered locations of all Indego bike stations is shown in the left-hand side of Figure 9. While Indego is a city-*owned* service, this figure shows it is not a city-*wide* service. Instead, Indego bike stations appear to be targeted at locations that have one or more of the following characteristics:

* A high volume of businesses and/or a base of large employers.
* Nearby high-volume Universities.
* A heavy population of millennial residents.

A picture containing text, map

Description automatically generatedA picture containing text, map

Description automatically generated

Figure 9- Philadelphia City Limits and the Location of Indego Bike Stations

The right-hand side of Figure 9 zooms in on the southern portion of Philadelphia that includes bike stations. This image shows stations marked by one of four colors. The four colors represent the 2018 usage volumes of the stations as described in Figure 10.

|  |  |
| --- | --- |
| **Color** | **Usage** |
| **Light blue** | Stations from which less than 5,000 trips originated. |
| **Medium blue** | Stations from which 5,000 to 9,999 trips originated. |
| **Orange** | Stations from which 10,000 to 14,999 trips originated. |
| **Red** | Stations from which 15,000 or more trips originated. |

Figure 10- Legend of Color Markers

The histogram, shown in Figure 11, shows the volume of 2018 trips taken from each starting station with the color stack representing the pass type used for trip. Although not labeled, each point along the x-axis represents one of Indego’s 130 stations.

A close up of a map

Description automatically generated

Figure 11- Frequency of Trips by Bike and Pass Type within Bike

Two stations are called out on the histogram with red arrow overlays. These stations are highlighted because they have a disproportionately high volume of Day Pass trips. As shown in Figure 12 these are the only two stations with Day Pass percentage volumes of over 50%.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Station | Day Pass | Indego30 | Indego365 | IndegoFlex | Walk-up | Day Pass % |
| 3049 | 2090 | 1593 | 219 | 37 | 134 | 51.31% |
| 3057 | 6682 | 3744 | 759 | 71 | 742 | 55.69% |

Figure 12- Pass Type Volumes for Stations 3049 and 3057

Why is the pass type distribution skewed for these two stations? Their locations may tell part of the story. The map provided in Figure 13 shows two white markers. Both stations are located along at the edge of one of Philadelphia’s two rivers. Station 3057 (left) is located on the bank of the Schuylkill River, not far from the Philadelphia Art Museum, and station 3049 (right) is located on the bank of the Delaware River in the heart of Philadelphia’s Penn’s Landing district. Both spots attract a lot of tourism and both provide scenic routes along the water’s edge. The single-trip Day Pass is a good option for a tourist in a scenic part of town wanting to enjoy the pleasant flow of the water and get some exercise.

A close up of a map

Description automatically generated

Figure 13- Location of Stations with More Than 50% Day Pass Usage

### What Routes are Being Taken?

A number of analyses were undertaken to determine the most heavily traveled routes, starting with a heatmap-based exploratory analysis, shown in Figure 14, with the x-axis representing individual start stations and the y-axis representing end stations. An unannotated version of the heatmap is shown on the left-hand side of the figure. Although the scale of the image combined with the volume of stations make pinpoint analysis difficult, a number of hot-points and symmetries emerge when using the artist’s trick of squinting to discover patterns. Some of these are annotated with arrows, circles and lines on the right-hand side of the figure.

There is a clear pattern on the bottom-left to top-right diagonal which is the line that represents x = y or the start station equal to the end station. In other words, this line represents round trips. While it was shown in Figure 7 that one-way trips far outnumber round trips, the heatmap shows that the 130 possible round-trip routes have notable ridership when compared to the remaining 16,670 one-way routes.

A close up of a map

Description automatically generatedA picture containing sky, water, man, skiing

Description automatically generated

Figure 14- Heatmap of Routes Between Stations

There are also some hot-point symmetries highlighted on the right with arrows. On the left, you will see hotter spots at each ends of these symmetries without the arrows. The symmetry of the arrows implies that these are two one-way routes taken between a set of two stations. Perhaps due to one-way ridership in the AM and ridership in the opposite direction in the PM.

Finally, the single hottest (darkest) point is circled on the right-hand side. It is also on the one-way trip diagonal. Round trips using this station are, in fact, Indego’s most heavily taken 2018 route.

### How Long Are Trips Taking?

Trips are charged in either one hour or one half-hour segments. Are people using the bikes for short destination-bound trip or are longer “joy rides” a common practice? Figure 15 provides a Pareto chart showing trip frequencies in half hour increment. (All trips lasting from one-minute to 30-minutes count as a one half-hour trip. Trips from 31-minutes to 60 minutes count as a two half-hour trip. Etc.) The graph is limited to trips taking five hours or less.[[11]](#footnote-11)

A screenshot of a cell phone

Description automatically generated

Figure 15- Pareto Chart of Trip Durations

Approximately 89% of all trips take one half-hour or less and 96% of all trips are complete within the first hour. With so many trips taking less than one half-hour, the granularity of this graph still leaves a lot of questions about trip durations. The histogram in Figure 16 shows the minute-by-minute frequencies of trips taking under one half-hour. As previously seen in Figure 7, the majority of trips are one-way, but we see a spike in very short round trips of one- to two-minutes. This may be due to riders redocking a bike after deciding to take a different bike or deciding to take an alternate means of transportation.

The mathematical mode of trips taking less than one-half hour is eight minutes with a log-style drop-off after the eight-minute mark. One explanation may be a high degree of destination-bound trips such as those taken from nearby residential areas to and from the business and university districts.

A screenshot of a cell phone

Description automatically generated

Figure 16- Drip Durations Under One Half Hour

## Bike Usage

Do bikes stay in a common area? Are all bikes used at a uniform rate? Figure 17 provides a histogram of bike usage in simple trip volumes. The Indego bike fleet consists of 1,466 bikes, each represented by a very thin line on the x-axis.

Clearly, bikes in the Indego bike fleet were not uniformly used in 2018. Given the limited bike information available, it is not clear how many of the 1,466 bikes used in 2018 were available for the full year or less. Were some removed from the fleet? Were some added? Were some taken out of service for a period due to damage and repair?

What is known, is that a small set of 10 electric bikes were introduced to the fleet in early November 2018. Electric bikes are represented by the thin red lines toward the right side of Figure 17.[[12]](#footnote-12) The use of these electric bikes is disproportionately heavy, given they were only available for approximately 10% of the year and not during the peak ridership months.

A picture containing text

Description automatically generated

Figure 17- Indego Bicycle Usage by Trip Volume[[13]](#footnote-13)

# Analyses

A series of analyses were made using multiple machine learning and classification algorithms. A summary of these analyses is shown in Figure 18. The best performing model for each analysis is reported in section 4 (*Results*).

|  |  |  |
| --- | --- | --- |
| Algorithm | Task | Target |
| Association Rule Mining | Feature Extraction | Learn higher-than expected features of Indego 2018 trips. |
| Feature Extraction | Determine most taken trip routes. |
| Clustering – K-means | Clustering | Can clusters be used to help a traveler in need find a nearby station with available bikes. |
| Clustering – Hierarchical | Clustering | What stations are in need of additional bikes? What stations are running low on available docks? |
| Decision Trees | Regression | Predict trip duration. |
| Random Forests | Classification | Predict destination station. |
| Support Vector Machines | Classification | Predict destination station. |
| Naïve Bayes | Classification | Predict destination station. |
| Text Mining | Frequency Analysis | Word cloud generation.[[14]](#footnote-14) |

Figure 18- List of Machine Learning Analyses Performed

## Association Rule Mining

Two models of association rule mining were employed. The first model sought to find interesting and non-obvious information that was contained in the Indego 2018 trip data. The second model targeted an understanding of the more frequent trip routes that were taken. In both cases the apriori() algorithm from R’s arules:: package was used, but with different parameter sets.

### Association Rule Mining to Determine Interesting, Non-obvious Information

The first association rule mining model was somewhat ad hoc in that there was no direct target, no specific questions whose answer was sought. Transactions were represented by individual trips. Each transaction consisted of the factor variables listed in Figure 19.

|  |  |  |
| --- | --- | --- |
| Variable | Source | Description |
| **bike\_type** | Indego data | Standard or electric bike. |
| **trip\_route\_category** | Indego data | One way or round trip. |
| **passholder\_type** | Indego data | Indego365, Day Pass, etc. |
| **start\_month** | Derived | Jan through Dec. Derived from the trip\_start date/time field of the trip. |
| **start\_day\_of\_week** | Derived | Sun through Sat. Derived from the trip\_start date/time field of the trip. |
| **start\_hour** | Derived | 12 through 11. Derived from the trip\_start date/time field of the trip. |
| **duration\_in\_half\_hours** | Derived | 1 through 49. Derived from the trip\_start date/time field of the trip. |
| **start\_station\_f** | Derived | Factorized version of the start station’s identifier. |
| **end\_station\_f** | Derived | Factorized version of the end station’s identifier. |

Figure 19- Variables Used in Untargeted Association Analysis

Prior exploratory data analysis revealed extremely skewed trip distributions for passholder\_type (Indego30 represents 72% of all trip records), duration\_in\_half\_hours (trips of 1 half hour or less represent 89% of all trip records), and bike\_type (standard bikes represent 99.7% of all trips). Therefore, rules containing any of the following transaction elements were disallowed from the models result set.

* bike\_type=standard
* duration\_in\_half\_hours=1
* passholder\_type=Indego30

To limit the dataset to a reasonable volume of rules with higher-than-expected support, the model’s statistical parameters were set at

* support = 0.01 – All rules had to occur in at least 1% of all trips. (6,365 trips).
* confidence = 0.1 – The right-hand side of a rule should occur at least 10% of the time for the given left-hand side.
* minlen = 2 – All rules must have at least two elements.

### Association Rule Mining to Determine Most Frequent Routes

The second use of association rules was targeted to a specific question. *What are the most frequent routes taken by Indego riders?* To obtain this answer each trip again constituted a single transaction. However, for this model each transaction contained only two elements, one for the start\_station and a second for the end\_station. The following parameters were passed to apriori().

* support = 0.00001 – Given the 16,900 possible start/end station permutations from 130 stations, a much lower support value was required.
* minlen = 2 – Assure all rules would have at least two elements.
* target = frequent itemsets
* appearance = list(lhs = *<start rules>*)) – Forces the start\_station elements to the left-hand side of a rule.

Note that no confidence parameter is specified and instead there is a new target parameter. The common value used for the target parameter is “rules.” Using “**frequent itemsets**” instead instructs apriori() to extract the item sets that appear together most frequently in the transaction list. This is essentially the same as sorting the rules in descending order of support.

Additionally, all rules of the form “start\_station=*<station id>”* were gathered into a single vector referenced as *<start rules>* in the parameter list. Only the list of items in *<start rules>* were permitted on the left-hand side of any rule. This, combined with min\_length=2 forces all resulting rules to be of the form

<start\_station=*<start station id>*> => <end\_station=*<end station id>*>

and when combined with the target=frequent itemsets parameter, the result is a list of the most taken routes, sorted in descending order of frequency.

## Clustering

As with association rule analysis, two separate forms of cluster analysis were performed with each analysis targeting a separate use case or business question. However, unlike the association rule analyses, both of which used the common arules::apriori() function, the clustering analyses undertaken use two separate algorithms. Additionally, the clustering analyses were performed on stations, not on trips.

### Finding Interesting Clusters With K-Means

K-means clustering was run to determine how a subset of variables could and would be clustered together and to determine if the clusters discovered had any business benefits. K-means analysis employed the kmeans() function of the R’s stats:: package and was applied to the subset of station variables shown in Figure 20.

| Variable | Source | Description |
| --- | --- | --- |
| **kioskId** | Indego data | Unique station identifier. |
| **latitude** | Indego data | Latitude location of the station. |
| **longitude** | Indego data | Longitude location of the station. |
| **bikesAvailable** | Indego data | Number of bikes currently available for rent at a station. |

Figure 20- Variables Used in K-mean Clustering Analysis

Using only location data (latitude and longitude) would be analogous to clustering points on a 2-dimenional plane. However, it was unclear what adding a third dimension of available bikes would produce. The final and most interesting model was created by setting the number of clusters parameter, K, to 5.

### Hierarchical Clustering

Initially, hierarchical cluster modeling was also initiated with no single business question in mind, so a variety of cluster models were run. In all cases clustering and cluster management was performed with the hcust() and cuttree() functions of R’s stats:: package. The model reported in the later results section (section 4.4) uses the variables described in Figure 21.

|  |  |  |
| --- | --- | --- |
| Variable | Source | Description |
| **kioskId** | Indego data | Unique station identifier. (Used for identification, not as a predictor variable.) |
| **bikesAvailable** | Indego data | Number of bikes currently available for rent at a station. |
| **docksAvailable** | Indego data | Number of docks currently open at a station for a bike return. |

Figure 21- Variables Used in Hierarchical Clustering

Using the predictor variables from Figure 21 and a k value of 3 for the cuttree() function, the originally *ad hoc* analyses were able to assist with a possible business focus about stations in need of bikes or docks at any given time. Also, to keep the model simple and the resulting dendrograms readable, a sample of 16 random stations was selected and only those stations were used for the model.

The final hierarchical clustering model, described in the Results section, was performed using manhattan distance computation and the complete method for center computation.

## Decision Trees – Predicting Trip Duration

Decision tree analysis was run to determine if the Indego service could predict the duration of a trip once a bike is checked out using the subset of trip variables shown in Figure 22.

| Variable | Source | Description |
| --- | --- | --- |
| **duration** | Indego data | Trip time in minutes. This is the target variable. |
| **start\_month** | Derived | Jan through Dec. Derived from the trip\_start date/time field of the trip. |
| **start\_day\_of\_week** | Derived | Sun through Sat. Derived from the trip\_start date/time field of the trip. |
| **start\_hour** | Derived | 12 through 11. Derived from the trip\_start date/time field of the trip. |
| **start\_lat** | Indego data | Latitude of originating station. |
| **start\_lon** | Indego data | Longitude of originating station. |
| **bike\_type** | Indego data | Standard or electric. |
| **passholder\_type** | Indego data | Indego365, Day Pass, etc. |

Figure 22- Variables Used in Decision Tree Analysis

Note that latitude and longitude were used in lieu of discrete start\_station identifier. This is because the use of a 130-item element factor variable resulted in unusable execution speeds. The use of continuous latitude and longitude variables allowed for more efficient computation without losing any information.

Also, to keep run-time to a reasonable level, a random sample of 1,000 trips without replacement was selected for tree construction. Decision tree analysis used the rpart() function of R’s rpart:: package. Since the decision tree is being used for regression-style prediction instead of classification, the method parameter was set to “anova.” Also, the minsplit parameter was set to 30, requiring nodes containing less than 30 cases to be terminal nodes.

## Random Forests –Predicting Destination Station

The random forest analysis is the first of three analyses attempting to predict the terminal stations (end\_station) given a near-common set of variables. Many R classification-based prediction machine learning and prediction algorithms will not work with larger factor-based target variables and those that do not provide a rejection at runtime often fail to complete in a usable amount of time. Therefore, for these analyses only trips whose source and destination were in the top 16 most-used stations were considered. This allowed the target to be limited to a factor variable containing no more than 16 unique elements and also trimmed the considered trip volume to 24,760 records.

The variables used for random forest analysis are shown in Figure 23.

|  |  |  |
| --- | --- | --- |
| Variable | Source | Description |
| end\_station\_f | Indego data[[15]](#footnote-15) | Factorized destination station. This is the target variable. |
| start\_time | Indego data | Date/time the trip started. |
| passholder\_type | Indego data | Indego365, Day Pass, etc. |
| start\_lat | Indego data | Latitude of originating station. |
| start\_lon | Indego data | Longitude of originating station. |
| duration | Indego data | How long the trip took, in minutes. |

Figure 23- Variables Used in Random Forest Analysis

Using duration may seem like cheating, but the rationale is that this type of model may be applied to a business problem such as searching for a bike that is currently on a trip. In this case, Indego staff would have access to the amount of time the trip has been in progress and thus, use of in-progress duration is available, valuable information.

## Support Vector Machines – Predicting Destination Station, Take 2

Support vector machine (SVM) analysis was the second analysis to attempt to predict the destination station. But, for SVM only continuous predictor variables were used, as shown in Figure 24. The ordinal passholder\_type variable was removed from consideration.

|  |  |  |
| --- | --- | --- |
| Variable | Source | Description |
| end\_station\_f | Indego data[[16]](#footnote-16) | Factorized destination station. This is the target variable. |
| start\_time | Indego data | Date/time the trip started. |
| start\_lat | Indego data | Latitude of originating station. |
| start\_lon | Indego data | Longitude of originating station. |
| duration | Indego data | How long the trip took, in minutes. |

Figure 24- Variables Used in SVM Analysis

The svm() function of R’s E1071:: package was used to perform all SVM analyses. Four separate kernels were attempted: linear, polynomial, sigmoid and radial. As will be discussed in section 4.7 (*Results of Support Vector Machines – Predicting Destination Station, Take 2*) the radial kernel performed best with the cost parameter set to a value of 15,000 and both the coef0 and gamma parameters set to 0. Additionally, all numeric value variables were scaled with the default scaling model[[17]](#footnote-17) prior to invoking the support vector function.

## Naïve Bayes – Predicting Destination Station, Take 3

Naïve Bayes is the third and final model that attempts to predict a destination station. The naïve Bayes experiments used the same variables that were used for SVM and was described in Figure 24. Naïve Bayes models were run using the naiveBayes() function of the E1071:: package. Models were run with both scaled and unscaled variable values with the laplace parameter set to 0.4 in the run with the highest accuracy. The best performer, whose result is reported in section 3.6 (*Results of Naïve Bayes – Predicting Destination Station, Take 3*) used scaled variables.

## Text Mining

Finally, a text mining exercise was performed. A corpus of four documents was created using text-only data from the four primary subsections of Indego’s FAQ page[[18]](#footnote-18): *Indego Basics*; *Managing Your Information*; *Passes*; and *Troubleshooting*. Each subsection containing from five to 10 separate questions and answers.

Text mining was performed using functions in R’s tm:: package. During translation of the corpus into a document-term matrix, the control list included options for removing punctuation, numbers and stop words and for stripping whitespace. To demonstrate the frequency of popular terms in the FAQs, a word cloud was generated using R’s wordcloud2:: package.

# Results

## Results of Association Rule Mining to Determine Interesting, Non-obvious Information

The first association rules model was designed to find non-obvious information contained in the Indego 2018 trip data. Given the criteria previously described, the apriori algorithm discovered 323 rules of length 2 and 266 rules of length 3 for a total of 589 rules. The 20 rules with the highest lift values are shown in Figure 25. Even among only these twenty rules there are at least three sets of rules that tell us patterns about Indego’s ridership.

A close up of a newspaper

Description automatically generated

Figure 25- Top 20 Rules (Sorted by Lift, Descending) of General Association Rule Mining

Rules [1] and [2] from Figure 25 show that Day Pass holders have a much higher than expected rate of riding round trips to and from station 3057. Recalling the information from Figure 12, station 3057 is the station located on the Schuylkill river, near the Philadelphia Art Museum. This adds credence to the theory that Day Pass tourists are tending to take a lot of joy rides along the bike paths at the bank of the river.

Next, rules [3] through [11], inclusive, when taken as a set, suggest another possible tourism trait. There is a two-to-three times higher than expected use of Day Passes for one-hour trips. Sometimes these trips are one-way, other times they are round trip.

Finally, rules [13] and [14] suggest that the weekends are associated with a higher-than-expected volume of Day Pass riders. Perhaps another trait of tourism or an indicator of residents exercising or in need of a simple means of transportation for running errands.

## Results of Association Rule Mining to Determine Most Frequent Routes

Having already explained how association rules and the apriori() function can be used to gather a pre-sorted list of the most frequently take routes, the 10 most frequent trips from the resulting list is shown in Figure 26.

A screenshot of a social media post

Description automatically generated

Figure 26- Top 10 Most Frequently Taken Indego Routes in 2018

The most frequently taken route is the round-trip course taken to/from station 3057, the station on the bank of the Schuylkill river that has appeared a few times already. This route was taken 2,743 times, an average of 7.5 times per day in 2018. Only two of the other nine most-taken routes are round-trip routies, number [7] taken to/from station 3053 and number [10] taken to/from station 3163.

Six of the remaining seven routes are mirrored pairs. The route between stations 3020 and 3032 is found in items [2] and [5]. Items [3] and [4] are for the route between stations 3054 and 3102. And items [6] and [9] are for the route between stations 3012 and 3020. Thus, station 20 appears in four of the top 10 routes.

The left-hand side of Figure 27 maps these top 10 routes.[[19]](#footnote-19) The white markers represent stations with round-trip routes. The blue markers show stations that are start/end points for one-way routes. The arrows show the direction of a route. Notice that three of the lined arrows have red and blue lines overlaid on one another. This can be seen by the butterfly shape that is made where the arrows meet. The right-hand side of the figure also shows how these top stations map back into the heatmap shown in Figure 14 with some of the top 10 station points annotated.

A close up of a map

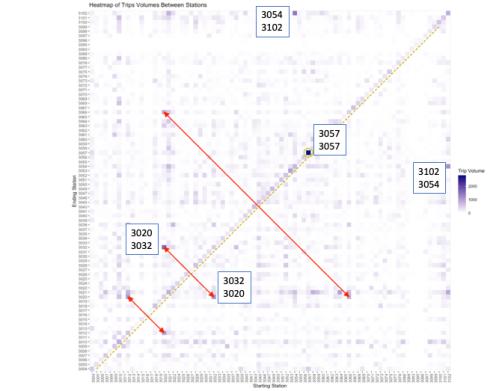
Description automatically generated

Figure 27- Top 10 Indego Trip Routes Taken in 2018

Four of the paths start and/or end in the University City district on the west side of the Schuylkill river. This area is home to the University of Pennsylvania (UPenn) and the Penn Health System and Drexel University. In addition to having large student bodies attending the universities, UPenn and the Penn Health System is the largest employer in the city and Drexel is the eighth largest employer.

The unmirrored one-way path terminates at the Comcast’s two high-rise towers, the two tallest buildings in Philadelphia. Comcast is the second largest employer in Philadelphia. The remaining two-way route, towards the top of the map, serves two museums, the Perelman annex of the Philadelphia Art Museum at the northern part of the route and the Rodin Museum at the southern part of the route.

## Result of Finding Interesting Clusters With K-Means

As mentioned in section 3.2.1, a value of K=5 was used to generate five clusters using latitude, longitude and the number of bikes available as input variables. The resulting clusters are shown on the map provided in Figure 28. Each cluster is represented by a separate color: blue, green, red, white and tan.

A close up of a map

Description automatically generated

Figure 28- Map of Five Clusters Using Latitude, Longitude and Bikes Available

There is a clear visual structure to the five clusters although many of them do not look visually clustered. For example, the white cluster is on the far left and far right with a gap in the middle. The red cluster is at the top and bottom. Why?

Figure 29 provides descriptions of the criteria that each cluster seems to follow. The provided cluster model appears to be based heavily on the centroids of all three variables provided to the model: latitude, longitude and number of bikes available.

| Cluster | Clustering Principle |
| --- | --- |
| Blue | Geographically centroid cluster with centroid volumes of available bikes. (Between 6 and 13 bikes available.) |
| Green | Geographically centroid cluster with extreme volumes of available bikes. (Less than 5 or more than 13 bikes available.) |
| Red | Cluster outside longitude centroid but still near the latitude centroid. (Centered between east and west but distributed north and south of the centroid.) |
| White | Cluster outside latitude centroid but still near the longitude centroid. (Centered between north and south but distributed east and west of the centroid.) |
| Tan | Cluster well outside the geographic centroid. (Far to the south.) |

Figure 29- Cluster Descriptions

While all five clusters are visualized in two dimensions, the addition of the third data dimension and the fact that each dimension has its own centroid, independent of the others, results in an sometimes-confusing visual model. To further demonstrate the clusters’ centroid models, Figure 30 provides a three-dimensional plot of the blue and green clusters. The x-dimension represents longitude, the y-dimension represents latitude and the z-dimension represents the number of bikes available. Notice the blue dots (stations) are clustered in the central portion of the z-axis which the green dots (stations) are at either extreme of the z-axis. However, the greens are still clustered around the centroid of the x- and y-axes, only the z-axis’ role is diminished.

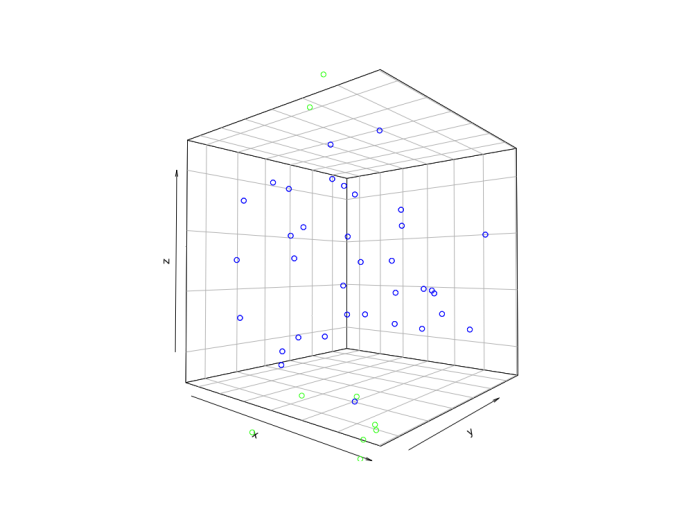


Figure 30- Three-dimensional Plot of the Blue and Green Clusters

Similar arguments can be made for the red and white clusters, centered around longitude and latitude, respectively.

## Results of Hierarchical Clustering

As mentioned earlier, hierarchical cluster modeling was initiated with no real plan or business question in mind. However, after experimenting with a few different variable combinations a model emerged that may have some business impact. A dendrogram of the resulting cluster hierarchy is shown in Figure 31.

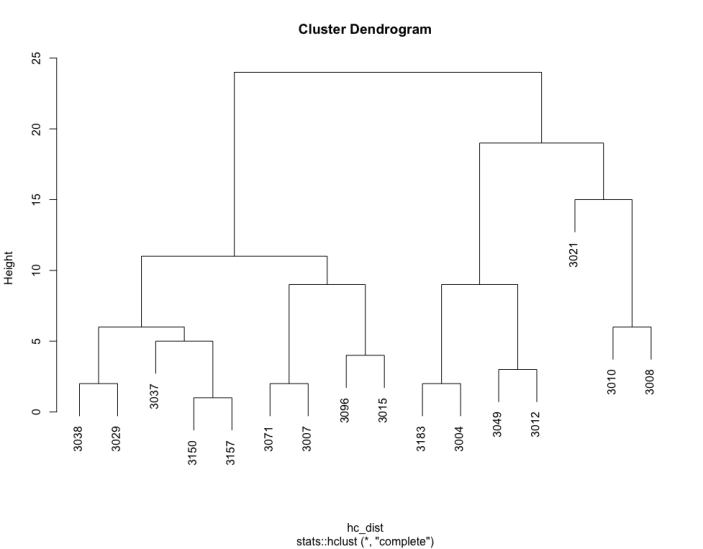


Figure 31- Dendrogram Produced from Hierarchical Clustering

This hierarchy was split into three clusters. Figure 32 shows the 16 stations, the cluster assignment, after cut, for each station, the bikes-available and docks-available dimensional values and three calculated fields: total docks (bikes available + docks available), delta (docks available – bikes available) and % bikes (bikes available / total docks).

| Station | Cluster | Bikes Available | Docks Available | Total Docks | Delta | % Bikes |
| --- | --- | --- | --- | --- | --- | --- |
| 3037 | 1 | 9 | 12 | 21 | 3 | 43% |
| 3150 | 1 | 7 | 9 | 16 | 2 | 44% |
| 3096 | 1 | 8 | 6 | 14 | -2 | 57% |
| 3038 | 1 | 10 | 9 | 19 | -1 | 53% |
| 3071 | 1 | 3 | 10 | 13 | 7 | 23% |
| 3007 | 1 | 5 | 18 | 23 | 13 | 22% |
| 3029 | 1 | 10 | 7 | 17 | -3 | 59% |
| 3015 | 1 | 5 | 5 | 10 | 0 | 50% |
| 3157 | 1 | 7 | 10 | 17 | 3 | 41% |
| 3010 | 2 | 17 | 4 | 21 | -13 | 81% |
| 3021 | 2 | 24 | 10 | 34 | -14 | 71% |
| 3008 | 2 | 13 | 6 | 19 | -7 | 68% |
| 3049 | 3 | 13 | 15 | 28 | 2 | 46% |
| 3183 | 3 | 16 | 12 | 28 | -4 | 57% |
| 3012 | 3 | 11 | 16 | 27 | 5 | 41% |
| 3004 | 3 | 17 | 13 | 30 | -4 | 57% |

Figure 32- Hierarchical Clustering Assignments

As can be observed from Figure 32, cluster one seems to include all the stations that are running low on bikes by volume – 10 bikes or less. Cluster two seems linked to a high percentage of bikes available which implied a lower percentage of docks are available. In other words, the stations in cluster 2 in may be filling up, limiting one’s availability to dock additional bikes. The third cluster seems to be those stations that are middling, with no immediate concerns.

## Results of Decision Trees – Predicting Trip Duration

As mentioned in section 3.3, decision tree analysis was performed using a random 1,000 item sample of the trip data with the goal of predicting trip duration. Given this sample and the variables described earlier, the decision tree shown in Figure 33 was created.

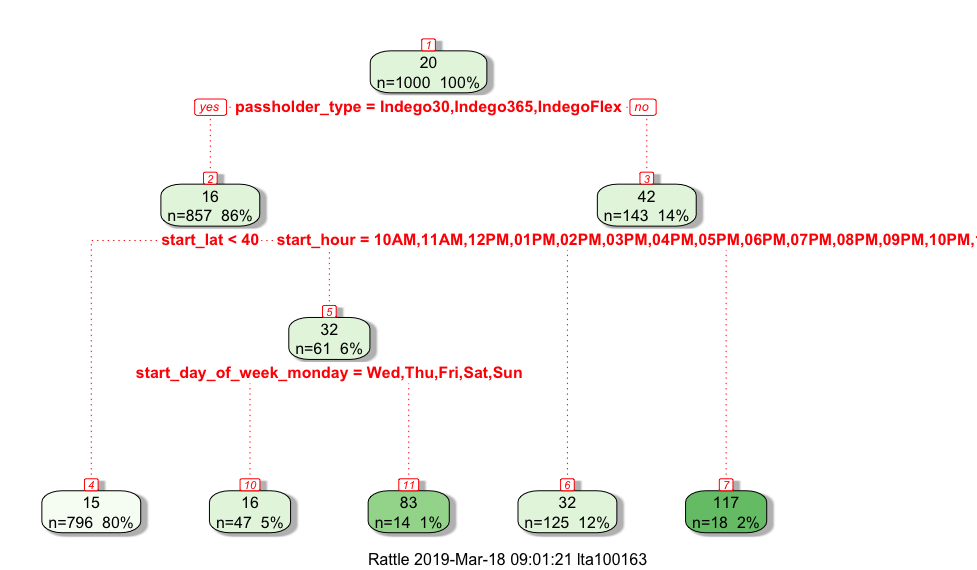


Figure 33- Trip Duration Prediction Decision Tree

The tree appears nicely structured, but how efficient is it? How does the tree perform on real data? The same data used to predict the tree was tested against the constructed tree and the predicted versus actual durations were plotted. (See Figure 34). The results do not look promising and are consistent with the tree’s large and wide-ranging mean-squared error (MSE) values spanning a low of 167 (node 5) to a high of 38,386 (node 11). An ideal plot on this graph would be a series of dotted steps from the bottom left and rising to the top right.

The five vertical lines of dots each represent one of the five terminal nodes of the decision tree. It has been previously demonstrated that the majority of trips are under 30 minutes and that within the 30-minute span, the trip duration mode was 8 minutes. However, the tree shown in Figure 33 has the shortest predicted trip durations of 14- and 15-minutes – a different of 75% of a large volume of trips.

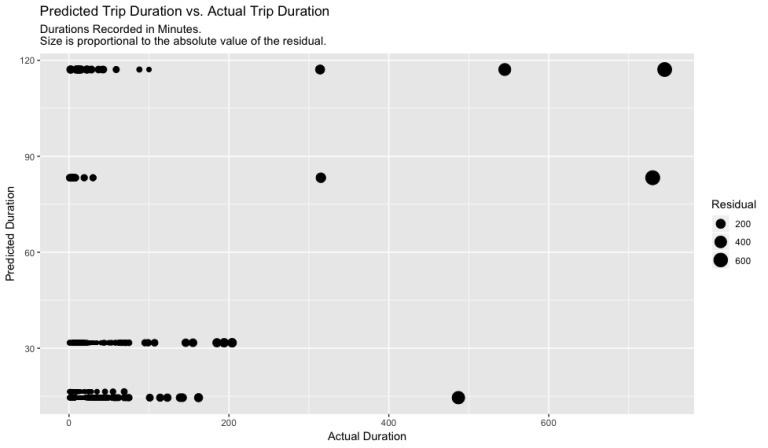


Figure 34- Predicted vs. Actual Duration from Decision Tree Analysis

An additional view of the performance of the algorithm is shown in Figure 35. Instead of plotting the predicted vs. actual, the residuals vs. the actuals are plotted. Even though most of the residuals seem to be on the low end of the graph, that low end actually spans 90 minutes of time.

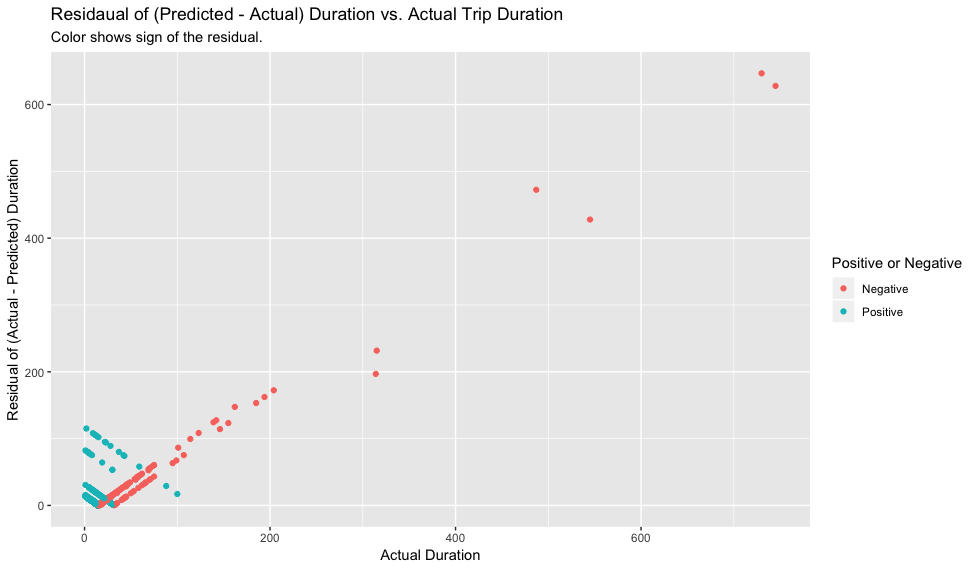


Figure 35- Decision Tree Prediction Residuals

## Results of Random Forests –Predicting Destination Station

The success of the random forest destination station prediction model is very much dependent on the use case it is applied to. A typical error plot (taken from the fifth fold of 10-folds cross-validation) is shown in Figure 36. The plot contains one error rate line for each of the 16 possible destination stations. The plot shows some stations with a 30% error rate and others with much higher error rates, at or above 70%.

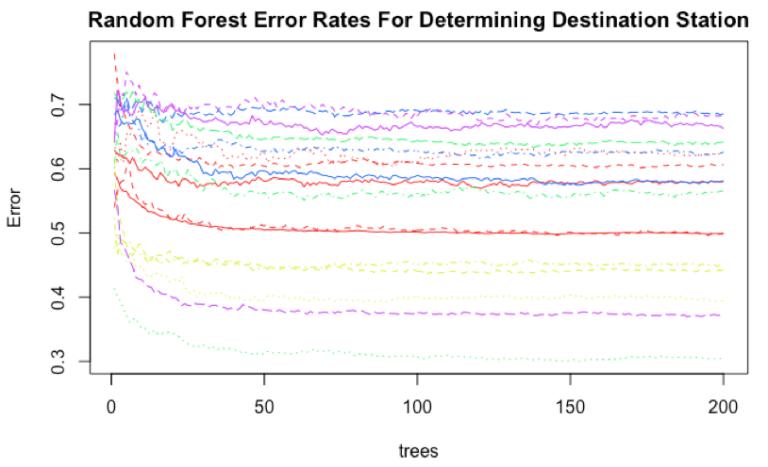


Figure 36- Typical Error Plot from Random Forest Analysis

The confusion matrix for the same fifth testing fold is shown in Figure 37. While the diagonal (the line of correct predictions does have darker color, as desired, the volume of shades tiles off the diagonal still represents too many incorrect predictions.

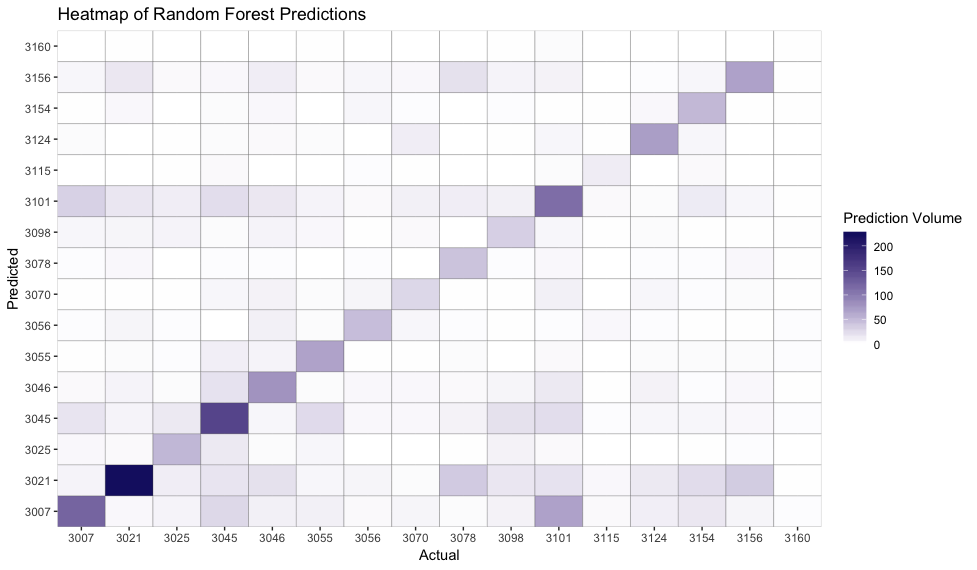


Figure 37- Confusion Matrix Heatmap from Random Forest Testing

Perhaps the best visualization for what use cases could benefit from this predictive model is shown in shown in Figure 38 – a boxplot of the error rate distributions for all sixteen (labeled) stations from all testing folds. In addition, the boxes are colored according to the station’s intake volume: higher intake is in red, lower in blue. This shows that the more heavily-used stations are more accurately predicted.

Given that the median values for six of the stations fall at or below 50% and four of those six are the four stations with the heaviest intake, this prediction may be beneficial for use cases such as law enforcement needing a set of probable insights as to where a rider may be destined. Or for inventory estimations at stations where higher degrees of intake is expected.

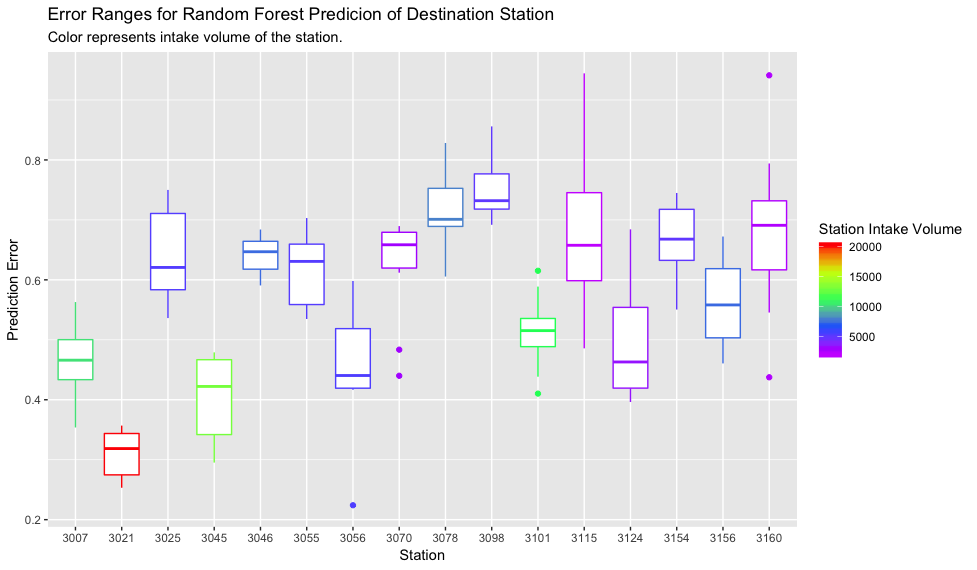


Figure 38- Boxplot of all Test Prediction Error Rates of Random Forest Testing

## Results of Support Vector Machines – Predicting Destination Station, Take 2

The SVM models performed worse that the random forest models. Of the four kernels used, the radial kernel performed best and is the only one reported here. For comparison with the random forest model an error rate boxplot for SVM is shown in Figure 39.

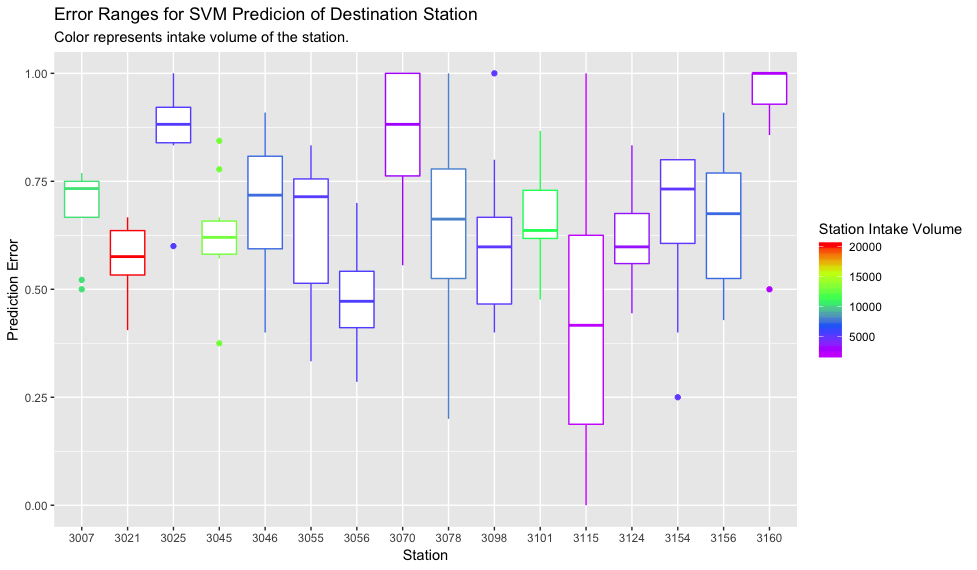


Figure 39- Boxplot of all Test Predictions Error Rates of Support Vector Machine Testing

Much higher error rates in SVM, even for this best model, can be seen immediately. Only two of the 16 stations had an error median below 50%. Additionally, SVM results do not seem to be biased by stations with a heavier intake. Perhaps one reason for this can be seen in Figure 40 where duration is shown as a function of start\_lat and start\_lon.[[20]](#footnote-20) (Reminder, the axis values represent the scaled values for start\_lat, start\_lon and duration, not the scaled values.) The previously-explained high frequency of trips under 30 minutes seems to cluster duration plotted points in a very small range of duration.

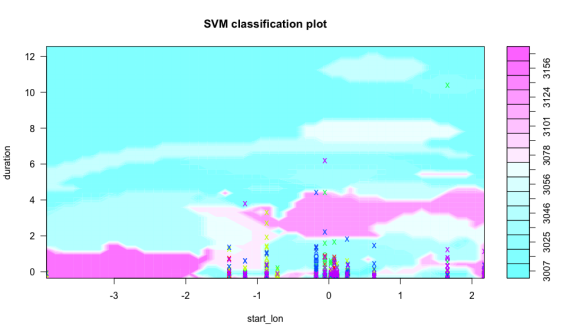
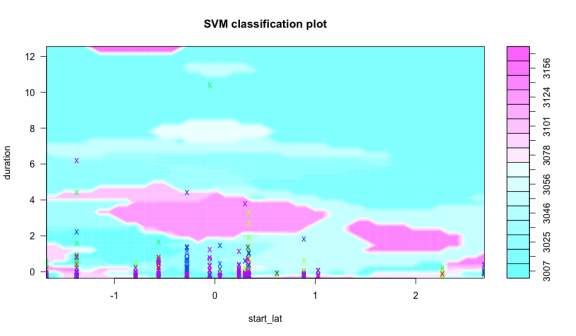


Figure 40- SVM Plots of Duration as a Function of Latitude and Longitude

Based on the results shown in these graphs, possible follow-on research should include scaling duration by some alternate model to try to spread the values out more or perhaps limiting the data sample to only those trips taking less than 30 minutes.

## Results of Naïve Bayes – Predicting Destination Station, Take 3

Naïve Bayes yielded the worst of all results for accuracy of prediction of destination station given the previously described predictor variables. Highest accuracy was reached using scaled variables and a laplace parameter value of 0.4.

The ten testing folds yielded accuracy results from a low of 15% to a high of 27% with a median accuracy of 18% and an accuracy mean of 19%. While these numbers do better than the guessing percentage of 6.25%, they are insufficient to drive any business action. Figure 41 charts the precision and recall values associated with each prediction point (station) from a cross-validation test that yielded an accuracy of 23.5%. While two of the recall values are between 70% and 80%, the general performance of naïve Bayes on this problem was poor.

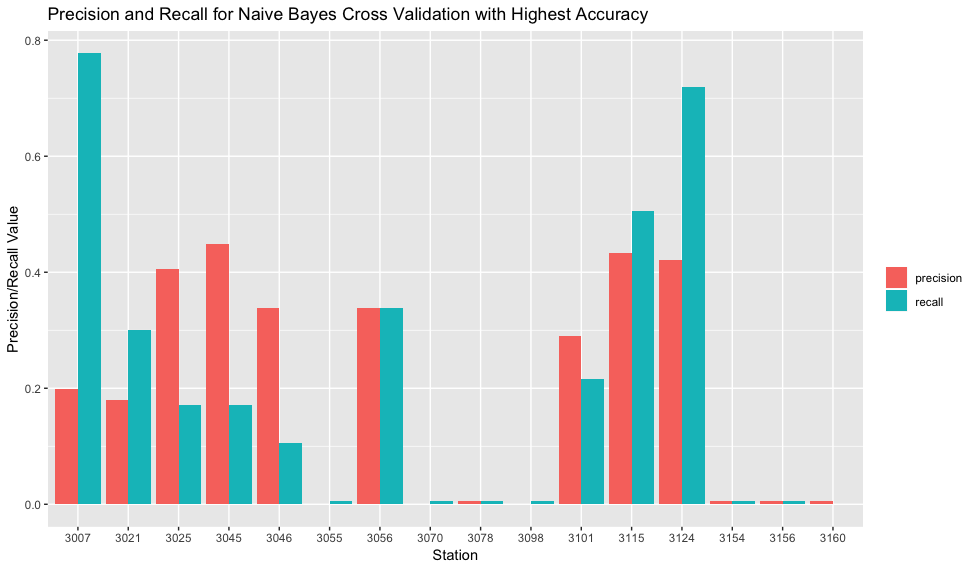


Figure 41- Precision and Recall Values for Naive Bayes

## Results of Text Mining

After mining the FAQ section of Indego’s website, the most frequently used terms were determined, and the top 150 most-frequent terms are displayed in the word cloud shown in Figure 42.



Figure 42- Word Cloud of Indego's FAQ Section

## A General Note on Results

At the start of section 3 (*Analyses*) Figure 18 provided a list of the analyses to be performed. As a bookend to that figure, Figure 43 provide a summary of the analysis findings.

|  |  |
| --- | --- |
| Algorithm | Summarized Result |
| Association Rule Mining | Exposed patterns of ridership, hypothesized to be associated with tourism. |
| Determined the most taken trip routes. |
| Clustering, K-means | Showed the ability to cluster stations according the geography and, bike availability for areas that have heavier populations of stations. |
| Clustering, Hierarchical | Showed the ability to cluster stations by those running low on bikes, those running low on available docks and those without concern. |
| Decision Trees | Attempted to predict trip duration. Showed some promise but not up to a production level of operation. |
| Random Forests | Attempted to answer if a destination station could be predicted for an in-progress trip. The random forest model showed results that may be practical for certain extreme use cases. More research is needed to provide more widely usable results. |
| Support Vector Machines |
| Naïve Bayes |
| Text Mining | A simple word cloud was generated. |

Figure 43- Summary of Results

# Conclusions

Freelance journalist Clive Thompson declared bike sharing to be today’s most exciting transportation model due to the transformation in bicycling that has been enabled by technology.[[21]](#footnote-21) With the technology-driven growth of bike sharing services, next-generation models enabling dockless bike sharing services are already being deployed. In addition to lowering costs due to the elimination of kiosks, a kiosk-free environment will allow bike sharing services to organically expand beyond the geographic limits currently served.

Indego is one of many technology-enabled bike share services growing rapidly in the United States. While Indego and other similar services currently operate within limited metropolitan areas, the ability to use a broader network of bike sharing services across state boundaries and even across countries could open up the bike sharing model in a way akin to how the ATM expanded personal banking services 35-years ago.

The additional of intelligence and connectivity into bikes can serve to increase bike share ridership. For example, smart bikes with built-in fitness application connectivity and heartrate monitoring could provide the bike sharing community with quantified-self information. Bike share services would then provide a fitness component for no additional cost and attract an even larger consumer audience.

Thanks to the availability of free and open data from Indego, patterns of customer usage, bike usage and routes have become evident. This allows for strong business models to be developed, optimizing customer needs while also increasing revenue.

1. Wikipedia. Bike-sharing System. 4 February 2019. <https://en.wikipedia.org/wiki/Bicycle-sharing_system>. [↑](#footnote-ref-1)
2. “*Disrupt*” is used in the sense of *business disruption*, defined by Clayton Christensen in *The Innovator’s Solution*. This is not a claim that bike share services are causing mass traffic jams although that may be the case. [↑](#footnote-ref-2)
3. National Association of City Transportation Officials. Bike Share in the U.S.: 2017. <https://nacto.org/bike-share-statistics-2017/>. [↑](#footnote-ref-3)
4. Reduced pricing is available to PA ACCESS card holders. PA ACCESS cards are available to residents requiring medical assistance and other low-income benefits. [↑](#footnote-ref-4)
5. Walk-up was phased out in 2018 and was replaced by the Day Pass option. [↑](#footnote-ref-5)
6. GeoJSON has become a fairly well-known format but GBFS is still more esoteric. Details on GBFS can be found at <https://github.com/NABSA/gbfs>. [↑](#footnote-ref-6)
7. Indego. Data – Trip Data – Data Processing. 2017. <https://www.rideindego.com/about/data/>. [↑](#footnote-ref-7)
8. Use of a negative latitude was clearly an error, unless Indego decided to host a kiosk in the southern Pacific Ocean, about 60 miles off the coast of Valdivia in southern Chile. [↑](#footnote-ref-8)
9. In many cases trip table latitudes and longitudes were not used. To ensure consistent station locations a trip’s station ID was joined back to the station table and the station table’s latitude and longitude were used instead. [↑](#footnote-ref-9)
10. These were most likely “virtual” stations that are set up for special events and removed at the end of the event. [↑](#footnote-ref-10)
11. If taken out to the 24-hour Indego data limit, the continual decrease of trip volumes continues but is meaningless for visualization. [↑](#footnote-ref-11)
12. The evenly spaced set of thin white lines in Figure 17 are rendering anomalies. [↑](#footnote-ref-12)
13. Sometimes nature and the regular human practices produce beautiful patterns, as is the case here with the elegant ogee curve of this bicycle usage histogram. [↑](#footnote-ref-13)
14. Hardly a true business case. [↑](#footnote-ref-14)
15. Both character and factorized version of the start and end stations were kept. Thus, an argument could be made that this factorized version (represented by the \_f in the name) is derived but it is considered direct source data for simplicity. [↑](#footnote-ref-15)
16. Since both character and factorized versions of the start and end station identifiers were stored in the station table an argument could be made that the factorized version (represented by the \_f in the name) is derived. [↑](#footnote-ref-16)
17. The default function for base::scale() is to divide a column its standard deviation, centered 0. [↑](#footnote-ref-17)
18. Indego. FAQ. 2017. <https://www.rideindego.com/faq/>. [↑](#footnote-ref-18)
19. Routes are defined here only by the start and end station pair. There is insufficient information to determine which streets and paths the bikes actually took. [↑](#footnote-ref-19)
20. The plots shown in Figure 34 are those produced by the default plot() function when a support vector machine model, a training data set and a function are passed as parameters. [↑](#footnote-ref-20)
21. Thompson, Clive. 2018 May 13. Wired. *The Vehicle of the Future Has Two Wheel, Handlebars, And Is a Bike*. <https://www.wired.com/story/vehicle-future-bike/>. [↑](#footnote-ref-21)