COMM 493 : ASSIGNMENT 2 – FORECASTING

Presented by: Nathan Saric Presented to: EcoCycle Dataset used: Bike Rentals Dataset	Weights BUSINESS – 40% TECHNICAL – 40% PROFESSIONALISM – 20%

Company Description

EcoCycle is a bike-sharing company founded in January 2018 that follows a sharing economy business model. The company oversees a large network of automated docking stations in major Ontario cities. It aims to provide an exceptionally smooth and effortless transport service by making bicycles available and accessible to all. The company's vision is to reduce traffic congestion and fuel consumption within cities by promoting a healthier and more economical alternative through their bike-sharing program. Sharing economy business models have become increasingly popular with brands such as Airbnb and Zipcar, providing lodging and transportation services, respectively, in an effort to maximize the efficiency of underused or idle assets. Sharing physical assets as services often allow for reduced costs since the emphasis is on accessibility as opposed to ownership. As a result, consumers are facing a paradigm shift by reevaluating their relationship with products and services, especially taking into consideration longer-term sustainability.

The key organizational values that EcoCycle continues to uphold since its beginning include dependability, simplicity, innovation, and safety. The company's primary value driver lies in its strategic vision to develop a circular economy that addresses both societal and environmental needs in its action plan. Additionally, the company's service offering closely aligns with consumer demands for affordable and flexible transportation alternatives paired with a growing interest in activities that seek to reduce one's carbon footprint.

EcoCycle's existing infrastructure allows users to borrow or rent a bicycle from an automated docking station and return the bicycle to any station belonging to the company. The stations are conveniently located at central hotspots around the city and are visible on all web mapping platforms such as Google Maps. The docking stations require no physical staff, as the stations are controlled entirely by computer, further reducing operational costs. Moreover, theft and vandalism are overcome by requiring a temporary deposit from the user if the bicycle is not returned within the subscription period. Lastly, the company faces two predominant user groups: casual users and registered users. The former represents the population that infrequently uses the bike-sharing program – most often one-time users – whereas the latter represents those with a membership that benefit from lower rental fees due to daily, or more frequent, use of the bicycles.

Problem Definition

Since the company's operations commenced in 2018, EcoCycle has gained significant traction across many communities through the steady rise in user engagement. Furthermore, the company has seamlessly integrated its infrastructure within cities and has received considerable support and funding from municipalities. In its first two years of operations, the company closely monitored activity levels across all docking stations in Ontario. The company hoped to ambitiously expand its network of stations to reach more profitable cities and capture new growth opportunities the following year. However, EcoCycle has recently undergone a period of uncertainty and instability due to the ongoing Coronavirus pandemic. On one hand, demand for outdoor activities, including cycling, has surged to unprecedented levels as people adapt to new forms of exercising which could potentially result in a shortage of public bicycles. On the other hand, the same surge in demand has incentivized people to purchase their own bicycles which eliminates the need for bike-sharing programs altogether. In either scenario, EcoCycle realizes that it is imperative to react quickly and strategically in order to remain competitive in such an evolving landscape. The company wishes to uphold its core values in a time of hesitancy and fluctuation by continuing to provide a safe and dependable service. In doing so, EcoCycle has reached out to better understand the demand for shared bikes and how the company should proceed in the coming months to regain its stability while trying to remain profitable.

Data Description and Transformation

In the following prediction report, we explore the data relating to daily bike rentals, between 2018 and 2019, in an attempt to provide insights and actionable strategies on how the company can more accurately forecast the demand for their service. We start by examining the dataset and commenting on any interesting elements in the data. Subsequently, we process and transform the data to train a predictor to ultimately forecast the demand for bike rentals for the upcoming month. Finally, we offer suggestions that allow the company to proactively boost and capture forecasted demand, with an emphasis on risk and EcoCycle's key value drivers.

The "*Bike Rentals*" dataset CSV file contains 15 numerical attributes in addition to the date attribute with 730 observations, indicating a daily observation over the span of two years. The first attribute represents the observation number, followed by 11 columns that depict various information about the time and weather for a given observation including the season, actual temperature (degrees Celsius), "feels-like" temperature (degrees Celsius), humidity, and wind speed (km/h). Additionally, the date is described in terms of the year, month, and weekday, along with whether or not the date was a holiday and/or a working day. The subsequent pair of attributes represents the number of casual and registered bike users. Finally, the 16th attribute is the daily count of rental bikes over the two-year period which will also be used as the target value we are interested in generating when forecasting into the future.

Before further analyzing the attributes, we first drop the observation number, year, month, holiday, and working day columns. The year and month are already given by the date attribute whereas the holiday attribute will be accounted for when training the predictor in *Amazon Forecast*. Next, we check for missing values in the dataset and find that the collected data has no missing values. We then create a PivotTable to explore the various columns. We observe that there were 3,290,845 total users between 2018 and 2019, of which approximately 19% were casual users and 81% were registered users. Moreover, the provided data dictionary illustrates that the season and weather attribute can take on one of four values (Appendix A). We see that about one-third of all bike rentals occurred in the summer with September 2019 experiencing the highest demand of 218,573 bike rentals, whereas January 2018 had a minimum demand of 38,189. Furthermore, plotting the "feels-like" temperature against the number of bike rentals highlights a positive correlation between these two attributes. Conversely, wind speed is negatively correlated with the number of bike rentals. These insights suggest that users are more likely to cycle in warmer and sunnier weather as opposed to cold, rainy, or snowy weather. We also note that despite the inherent fluctuation in demand from season to season (or month to month), the proportionality of the weekday attribute is consistent, such that any given weekday roughly accounts for one-seventh of the total demand over the two years.

We now transform and structure the data so that it is compatible with the *Amazon Forecast Environment*. We add a new column, *item_id*, which describes a unique identifier for the item we wish to predict and assign the value "*BIKE*" to each row for simplicity. Note that if EcoCycle were to also provide another form of transportation, this would be included as a separate *item_id*. Next, we relabel the *count* attribute to *demand*, and the *date* attribute to *timestamp,* to match with *Amazon Forecasts'* naming convention. Finally, we must clean the date column since the raw data is inconsistent. Converting the date to text and then using various Excel text functions (*LEFT*, *MID*, and *RIGHT*) to parse the date, we are able to reconstruct a proper date value. Note that it is important to format the date as "yyyy-mm-dd" in order for it to be compatible with *Amazon Forecast*.

Next, we duplicate the data into two individual CSV files, to be later supplied to the machine learning model. The first file will be the target time-series dataset which includes the *item_id*, the *timestamp*, and the *demand* attributes. Additionally, we may choose to include other attributes, however, the attribute which we want to generate a forecast for must be included. The supplemental attributes will become extra forecasting dimensions when training the predictor. The second file will be the related time-series dataset which will include all of the attributes except for *demand*. Supplying related data in addition to historical data can help increase the accuracy of the forecast by providing more context of various conditions that may have affected the demand. Lastly, although not included, EcoCycle may choose to supply the model with item metadata, a third dataset that can help refine forecasts by providing static information relating to the *item_id* such as the colour or model of the bicycles.

Data Import

Prior to importing the data into *Amazon Forecast,* we first need to create an *Amazon S3 Bucket* to store the two time-series datasets. After successfully uploading both CSV files to the bucket, we can now create a new dataset group in *Amazon Forecast*. Datasets are required to train predictors, which are then used to generate forecasts. We configure the forecasting domain to be "Retail" as this predefined domain best matches the context of our problem which is to forecast demand. We set the frequency of our data to be 1 day time intervals and then use the schema builder to replicate the structure of the target time-series dataset by specifying each attribute name and type in the same order as they appear in the CSV file. Recall that the timestamp format must be set to "yyyy-MM-dd". Next, we import the data by specifying the data location which is the path to the corresponding target time-series dataset file in the previously generated *Amazon S3 Bucket*. Lastly, we navigate to the Identity and Access Management (IAM) dashboard to obtain our custom IAM role ARN and supply this value to provide the dataset group with permission to access and read the contents of our *Amazon S3 Bucket*.

We repeat the above steps to import the related time-series dataset. Additionally, if the company wanted to provide item metadata, these steps would also be followed to import the third and final dataset. Again, the purpose of including related time-series data and item metadata is to provide additional dimensions of context which allows for a more comprehensive forecast from better predictions.

Predictor Training

Having properly imported the necessary datasets into *Amazon Forecast* we proceed by training a predictor. *Amazon Forecast* employs a custom model with underlying infrastructure to train on the supplied datasets. At this stage, we set the forecasting frequency to be one day time intervals, once more, and specify a forecast horizon. This value identifies how far into the future to predict the data at the given forecast frequency. Currently, EcoCycle recognizes that the demand for their service is volatile and has requested a forecast horizon of 31 days – forecasting the demand until January 31st, 2020. It is recommended for EcoCycle to produce monthly forecasts with a daily frequency, rather than weekly or yearly forecasts, as this will minimize variation from shorter-term forecasts while best accounting for seasonality in the time-series data.

Although not incorporated in this particular forecast, EcoCycle may wish to input multiple forecast keys to be used in training. These dimensions correspond to the columns in the target time-series dataset. *Amazon Web Services* documentation suggests that a *location* attribute may be a useful dimension in such a forecasting domain, which EcoCycle should consider collecting in the coming months. We then use the default forecast quantiles (0.10, 0.50, and 0.90) to create forecasts and evaluate predictors. Similarly, we use the default

optimization metric which lets *Amazon Forecast* create an optimized predictor by selecting models with the lowest average losses over the specified forecast types. Lastly, we include holidays in the predictor training to improve forecast accuracy and set "Canada" as the input country. Note that the predictor is capable of producing an explainability report which attempts to illustrate the impact of each attribute on the target, however, experimenting with this resulted in all attributes having an impact score of zero (Appendix B).

Once complete, the trained predictor will yield various accuracy metrics which can be exported for inspection. In this instance, the root-mean-squared error (RMSE) was approximately 1,598 which indicates a fairly poor fit of the model to the actual data (Appendix C). The high absolute error is further reflected in the weighted absolute percentage error (WAPE), however, the mean absolute percentage error (MAPE) of 0.8157 is the preferred metric in this domain as it uses an unweighted average, which emphasizes the forecasting error on all items equally, regardless of their demands. Furthermore, the MAPE penalizes under-forecasting and over-forecasting equally. Thus, it is imperative for EcoCycle to fully comprehend the costs of under- and over-forecasting in the bicycle-sharing industry because the significance of the accuracy metrics should lessen if the difference in costs is not negligible. For example, the cost of under-forecasting may lead to a decrease in revenue and customer satisfaction meanwhile the cost of over-forecasting results in an excess of unused bicycles and, in turn, operational inefficiencies.

Forecast

After creating an *Amazon Forecast* predictor, we can finally generate a forecast based on the predictor we have trained. We specify the forecast quantiles for probabilistic forecasts to include the three previous quantiles (0.10, 0.50, and 0.90) as well as including *mean* and 0.99. Once the forecast is created, we have the option to query the forecast along with exporting the forecasted data. We choose to create a forecast export and supply our custom IAM role ARN, once more, to give permission to access and write the contents of the export to our *Amazon S3 Bucket*, supplied as the export location.

At this stage, *Amazon Forecast* will display an embedded line chart that displays the complete range of the forecast, by default, which can be queried by specifying filter criteria. Instead, we have created an interactive dashboard in Excel, to accompany this prediction report, that replicates the functionality of *Amazon Forecasts'* line chart in addition to other key visuals. To elaborate, the graphical representation shows the historical data in addition to the forecasted data. The forecasted region will display five sets of data that correspond to the five quantiles previously specified. Therefore, each data point represents the forecasted demand for bicycle rentals for a given date. The provided menu can be used to filter the data for a given time range, in the past and/or future and the dashboard will present the relevant information.

Returning back to the discussion of risk and EcoCycle's value drivers, the uncertainty associated with forecasts, in comparison to the target result, is expressed in prediction quantiles. For example, showing the complete range of forecasted data in the dashboard, one can interpret that the demand for bicycle rentals on January 1st, 2020 will be roughly 3,991 bicycles at the 90% quantile (P90 Forecast) or 2,957 bicycles at the 50% quantile (P50 Forecast). In other words, 90% of the time, the true demand for bicycles will be less than 3,991 whereas 50% of the time, the true demand for bicycles will be less than 2,957. Therefore, EcoCycle must factor in the risk relating to its strategic vision to address both societal and environmental needs in its action plan. The company must identify its degree of risk tolerance and interpret the forecast accordingly. Should the company choose to have a higher degree of risk tolerance and assume a higher demand for any given date, the company will need to produce more bicycles and/or modify the automatic docking stations to be able to support a larger number of bicycles in day-to-day operations. On the other hand, assuming a lower degree of risk tolerance, the company may encounter a shortage in bicycles and may forfeit customers to a competing company. In either case, the discrepancy between supply and demand should be minimized in order to regain stability and allow EcoCycle to resume operations at optimal levels of efficiency.

Sources

The following *Amazon Web Services* documentation were referenced when initializing, configuring, and troubleshooting the *Amazon Forecast Environment:*

https://docs.aws.amazon.com/forecast/latest/dg/what-is-forecast.html

https://aws.amazon.com/blogs/machine-learning/tailor-and-prepare-your-data-for-amazon-forecast/

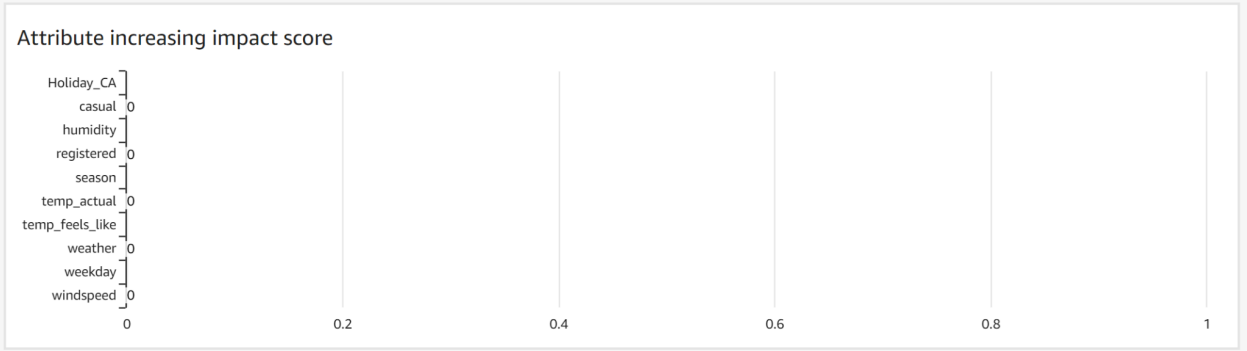
https://aws.amazon.com/blogs/machine-learning/measuring-forecast-model-accuracy-to-optimize-your-business-objectives-with-amazon-forecast/

Appendix

Appendix A: Data Dictionary

item_id	string	"BIKE"
timestamp	timestamp	2018-01-01 to 2019-12-31
season	integer	1: Winter, 2: Spring, 3: Summer, 4: Fall
weekday	integer	1: Sunday, 2: Monday, 3: Tuesday, 4: Wednesday, 5: Thursday, 6: Friday, 7: Saturday
weather	integer	1: Clear/Sunny, 2: Mist/Cloudy, 3: Light Snow/Rain, 4: Heavy Snow/Rain
temp_actual	float	-100 ℃ to +100 ℃
temp_feels_like	float	-100 ℃ to +100 ℃
humidity	float	0 to 100
windspeed	float	0 to 100
casual	integer	Any value greater than or equal to 0
registered	integer	Any value greater than or equal to 0
demand	float	Any value greater than or equal to 0

Appendix B: Explainability Results



Appendix C: Predictor Accuracy Metrics

Weighted Quantile Loss [0.10]	0.2046
Weighted Quantile Loss [0.50]	0.3101
Weighted Quantile Loss [0.90]	0.1187
Average Weighted Quantile Loss	0.2111
Root Mean Squared Error	1,597.9711
Weighted Absolute Percentage Error	37,831.3654
Mean Absolute Percentage Error	0.8157
Mean Absolute Scaled Error	1.4918