Rider Volume

Nicholas Morris

Machine Learning

Wrangling → Feature Engineering → Modeling

Daily Volume

Adding previous days of rider volume to the data by route.

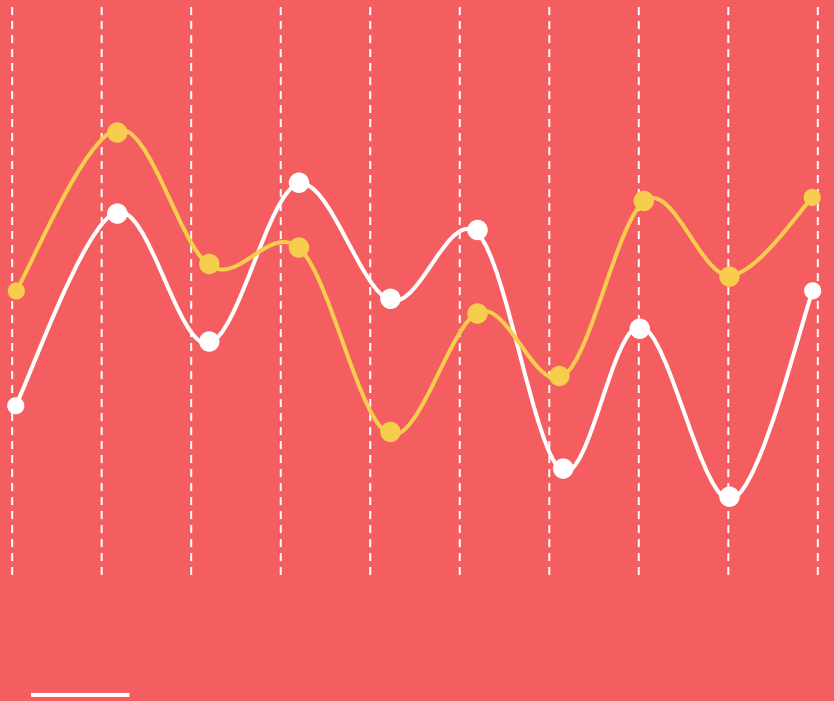
Additional Data

Adding economic trends such as unemployment and CPI. Converting timestamp components and route to binary data points. Attempting feature transformations.

Predictions

Training linear regression, XGBoost, and deep learning neural network models. Evaluating performance. Computing feature drift to signal retraining.

Wrangling



Dataset

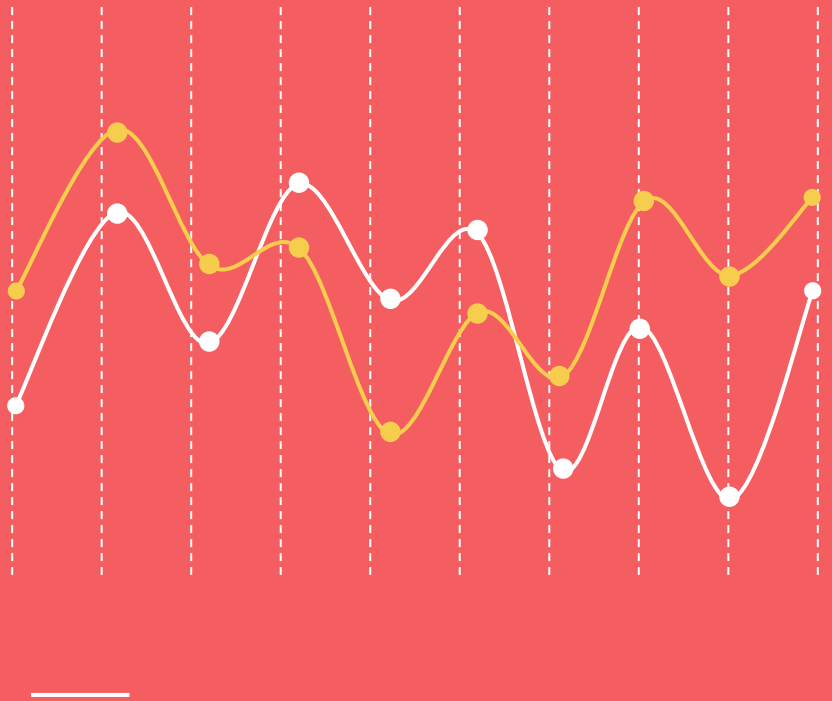
route	date	daytype	rides
3	01/01/2001	U	7354
4	01/01/2001	U	9288

Below is the first two entries in the data. There are 980,744 total entries and 4 columns. The target we are predicting is rides. [Link to the dataset and code]

Previous Days Of Rider Volume

The last four days of rider volume per route and day type were added to the data.

Feature Engineering



Binary Data

Route, day type, and timestamp components were converted to binary data points.

Economic Data

The NASDAQ closing price, Unemployment rate, CPI, PPI, GDP, GDI, and Federal Funds Rate were pulled from FRED (Federal Reserve Economic Data).

Atwood Numbers

An Atwood Number is a calculation that shows the relative change between two variables. The formula for two variables x and y is:
(x - y) / (x + y)

This calculation was done on all pairs of non-binary variables; but did not improve model performance, so, it was left out of the final model.

Binning

Binning is when a non-binary variable is grouped into histogram bins, and represented as binary variables.

Binning did not improve model performance, so, it was left out of the final model.

Reciprocals

A reciprocal is when a non-binary variable x is calculated as 1 / x.

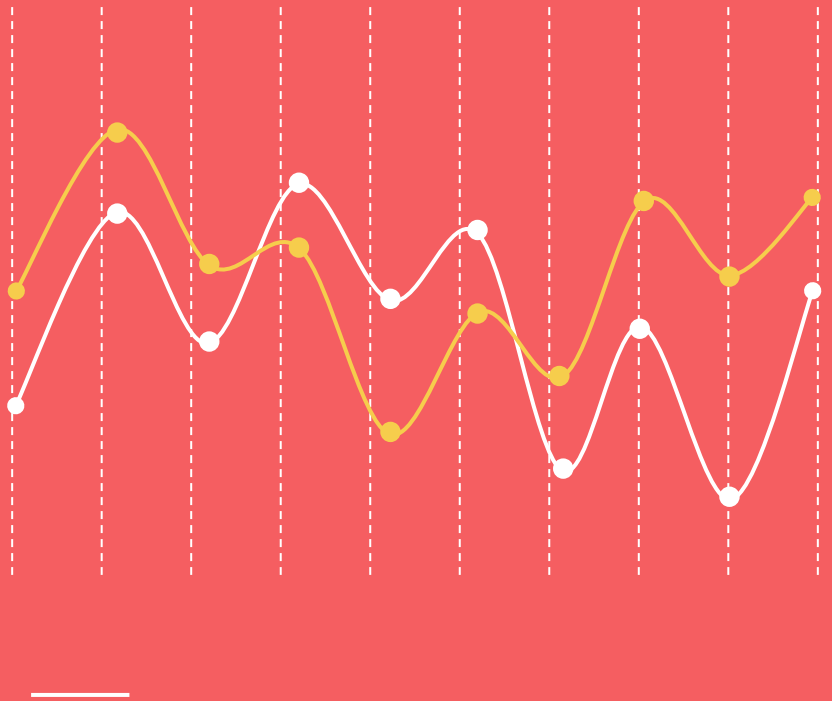
Reciprocals did not improve model performance, so, it was left out of the final model.

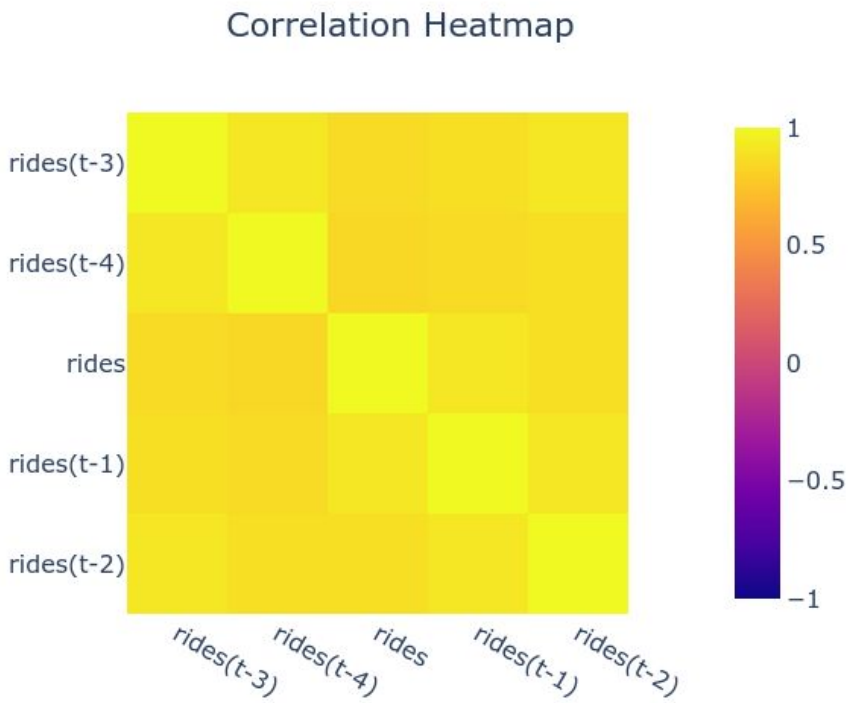
Interactions

An interaction is when two variables x and y are calculated as x * y. Reciprocals were fed into this calculation to generate x / y as well.

Interactions did not improve model performance, so, it was left out of the final model.

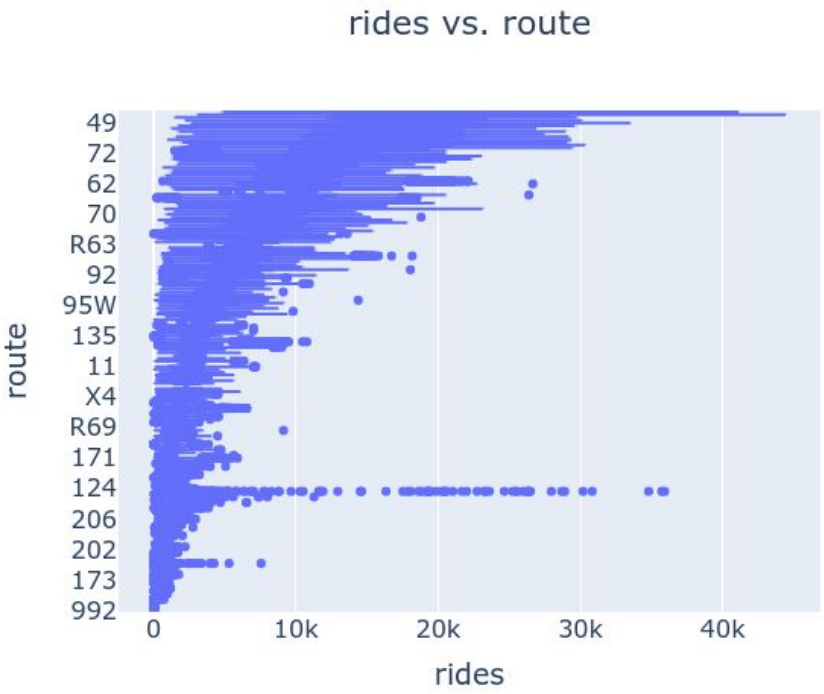
Data Exploration



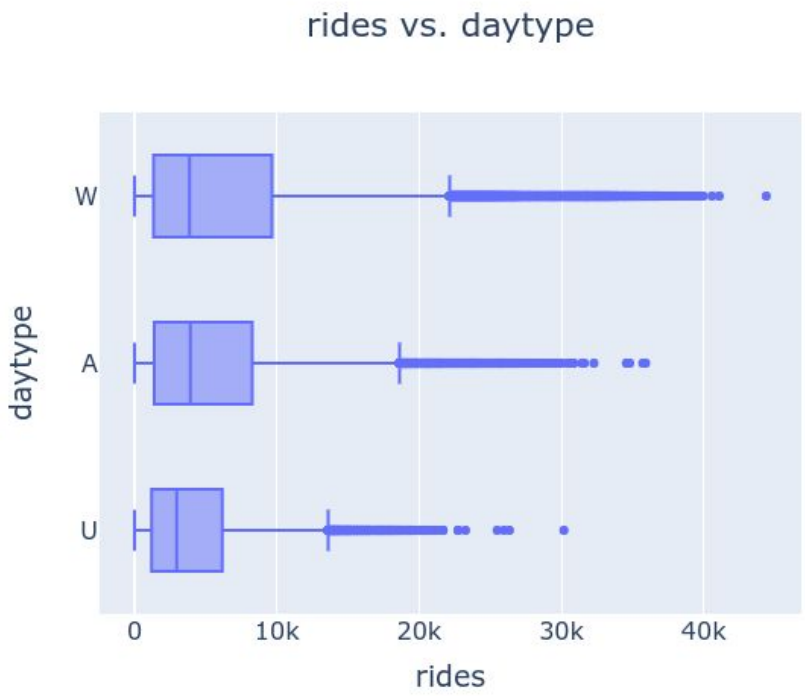


There is a strong correlation between rider volume and previous days of rider volume.

There is a positive trend between rider volume and the previous day of rider volume, which shows that when rider volume is low or high so is the previous day. There are some outliers on the edges though.

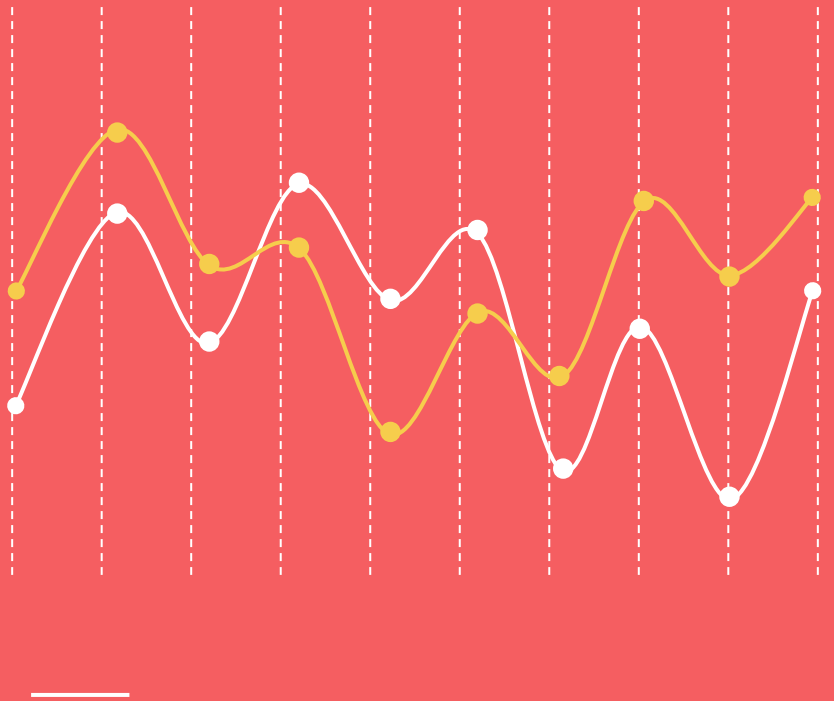


Rider volume varies by route with route 49 having much higher volumes on average compared to route 992.



Rider volume is slightly different according to day type. Day W has a median volume of 3860 riders and day U has a median volume of 2958 riders.

Modeling



Model Parameters

Linear Regression

Library: scikit-learn
Length Of Path: 1e-9
Number Of Alphas: 16
Cross Validation Folds: 3
Tolerance: 1e-4
Max Iterations: 500

XGBoost

Library: xgboost
Boosting Rounds: 100
Learning Rate:
0.001, 0.01, 0.1
Max Depth:
5, 7, 10, 14, 18
Min Child Weight: 1
Column Sampling: 0.8
Row Sampling: 0.8
Cross Validation Folds: 3

Neural Network

Library: Tensorflow
Epochs: 500
Learning Rate:
0.0001, 0.001, 0.01
Batch Size: 16
Layers: 10
Nodes Per Layer:
32, 64, 128, 256, 512
Solver: Adam
Cross Validation Folds: 3

Model Comparison

Linear Regression

R2: 0.92
RMSE: 1786
In Control: 96.26%

Model Indicators:
1. rides(t-1)
2. route_79
3. route_9
4. route_66
5. route_4

XGBoost

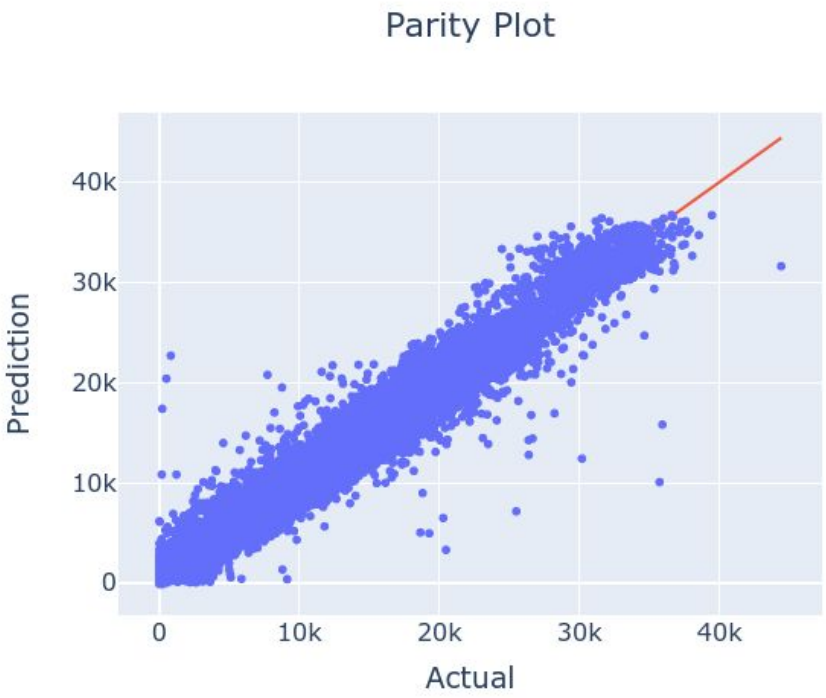
R2: 0.98
RMSE: 855
In Control: 95.76%

Model Indicators:
1. date_year_2020
2. rides(t-1)
3. route_79
4. date_gdp
5. route_9

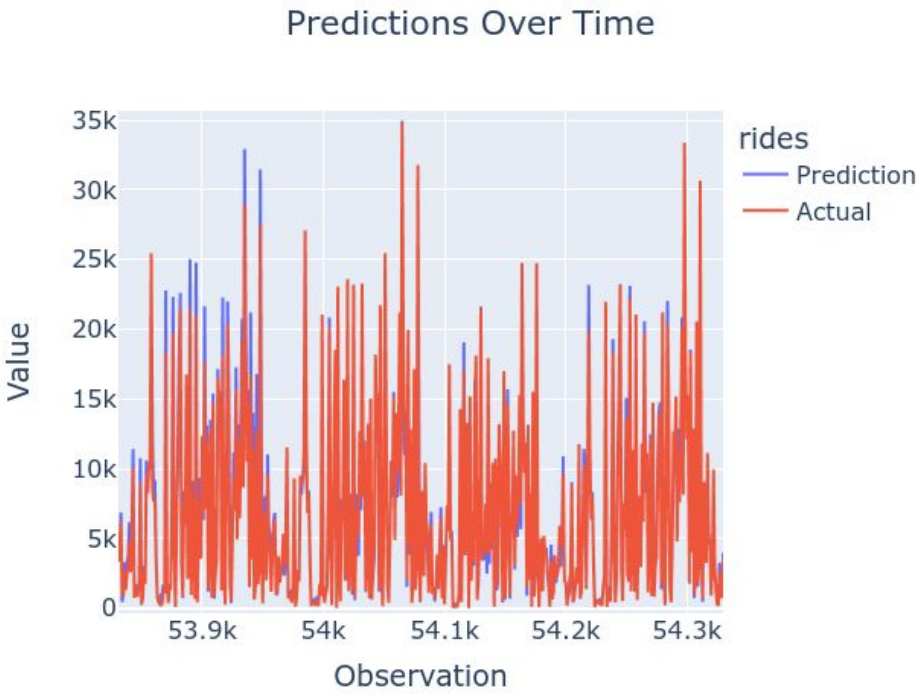
Neural Network

R2: DNF
RMSE: DNF
In Control: DNF

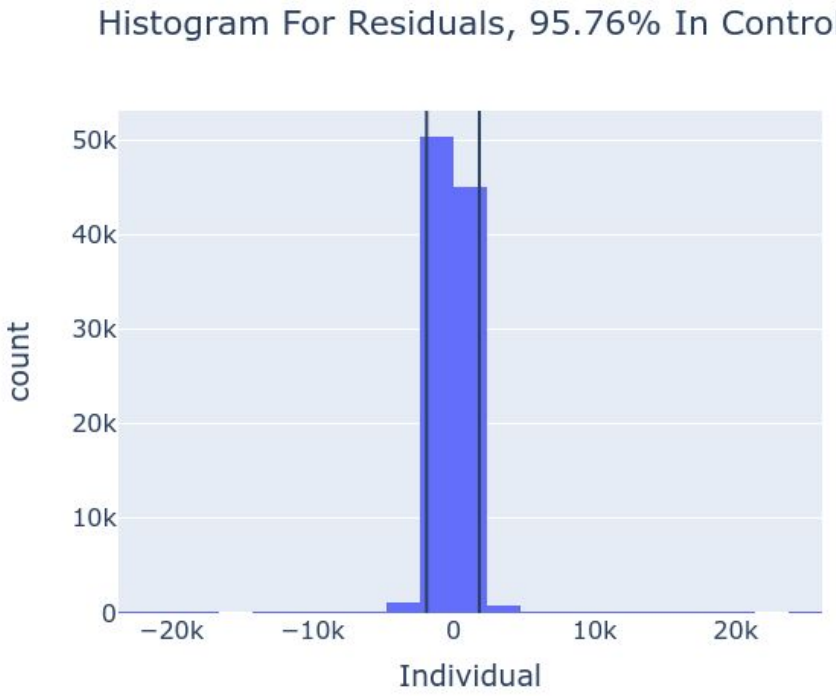
Model Indicators:
DNF



These predictions come from the XGBoost model. These predictions are done on 20% of the data that the model did not see during training. The predictions are centered on the red line (perfect predictions). There are scattered outliers over and under-predicting rider volume.

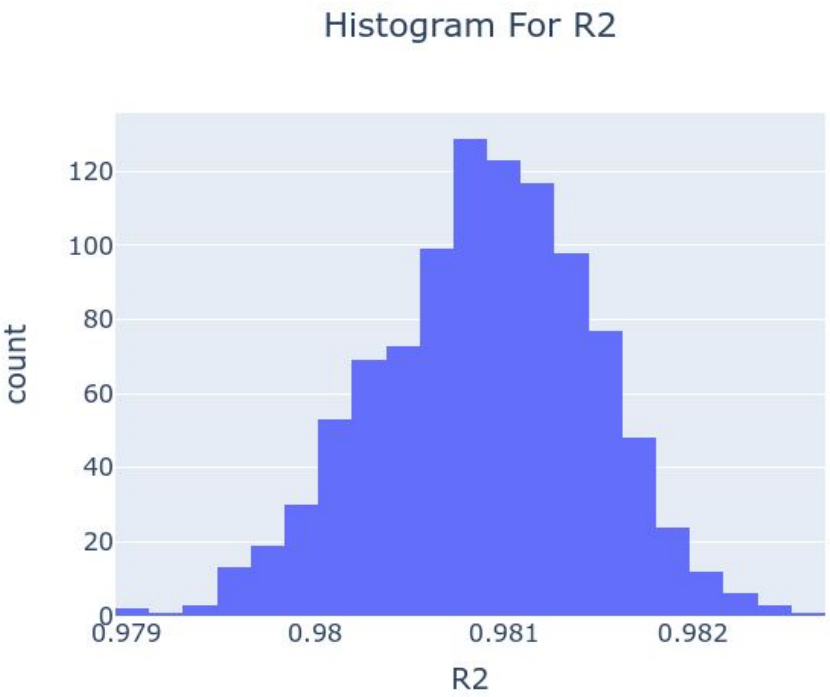


Here's a snapshot of the predictions over time. We can see the the blue predictions follow the actual values well. There are some actual values that jump up past the predictions. We can see there is some seasonality in the data as well.



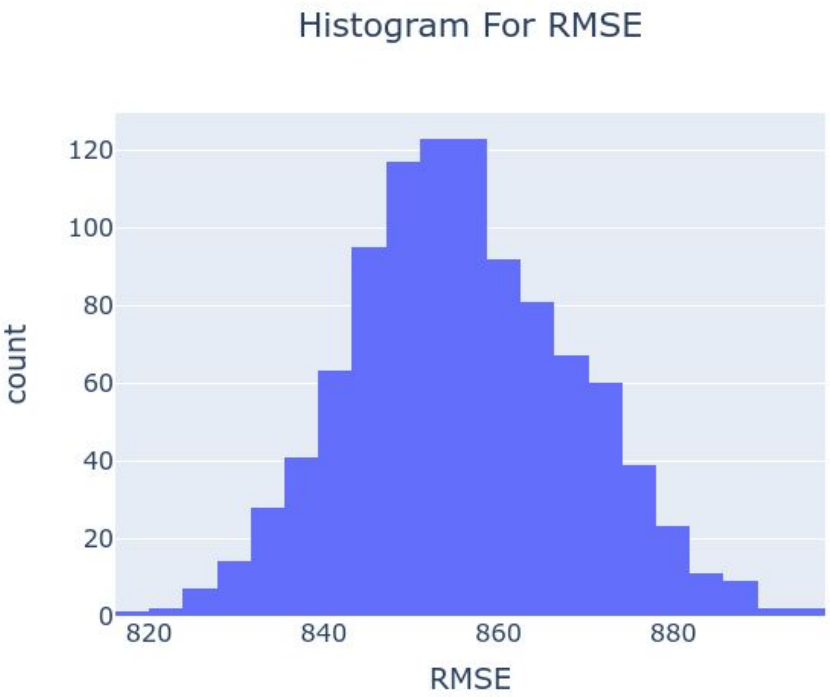
The residuals are prediction error = actual - predicted.

The residuals have a tight bell shape, which is good, and they are centered on zero. Control limits were computed on the residuals and we can see that the prediction error is mostly under control. We can see small, long tails which indicates a small tendency to over and under-predict rider volume.



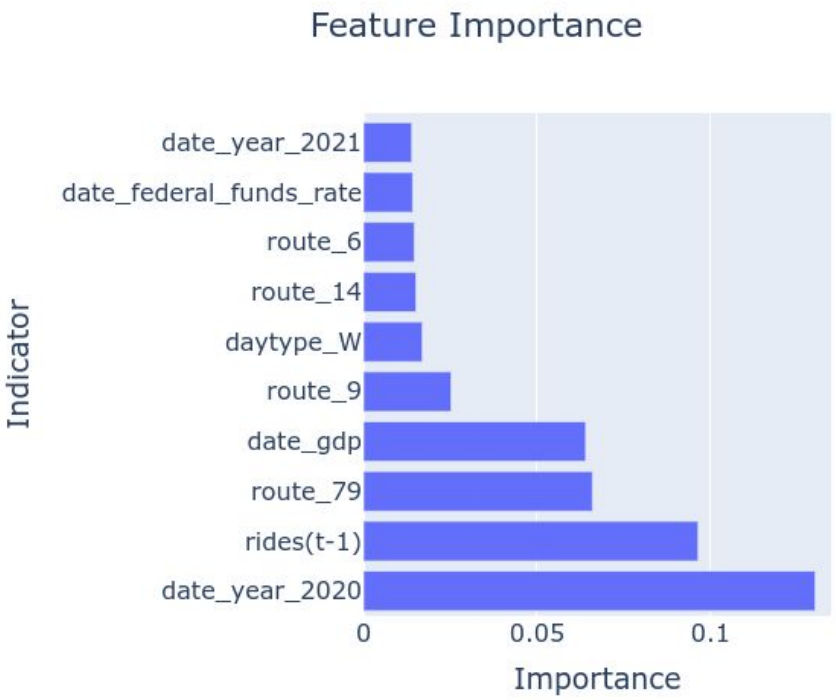
The prediction error was resampled 1000 times at a 50% sampling rate with replacement. Then R2 was computed on each sample to get a distribution.

R2 has a tight range between 0.979 and 0.982, which is good. R2 has a bell shape, which is good, and a slight skew to the left.

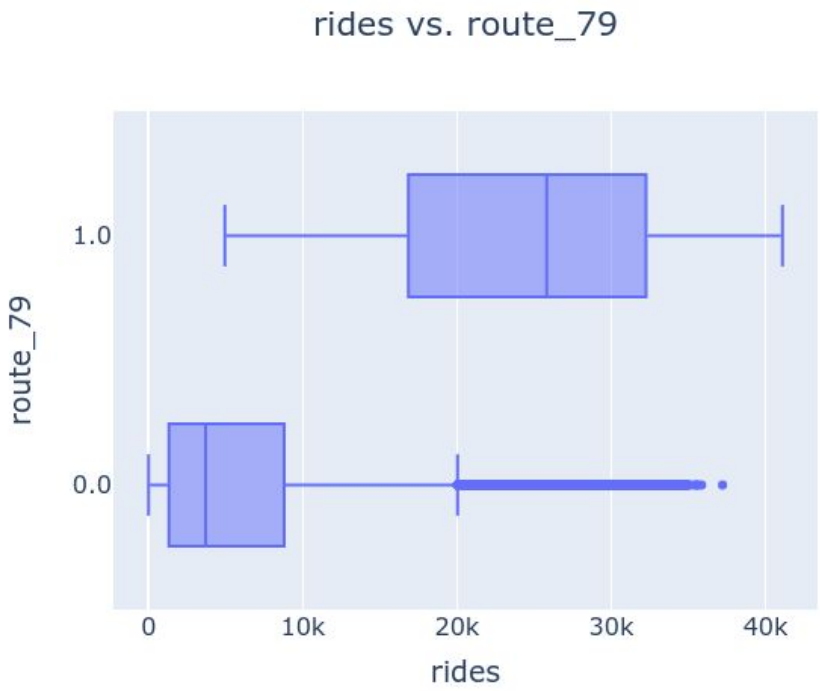


The prediction error was resampled as previously mentioned to get a distribution for RMSE.

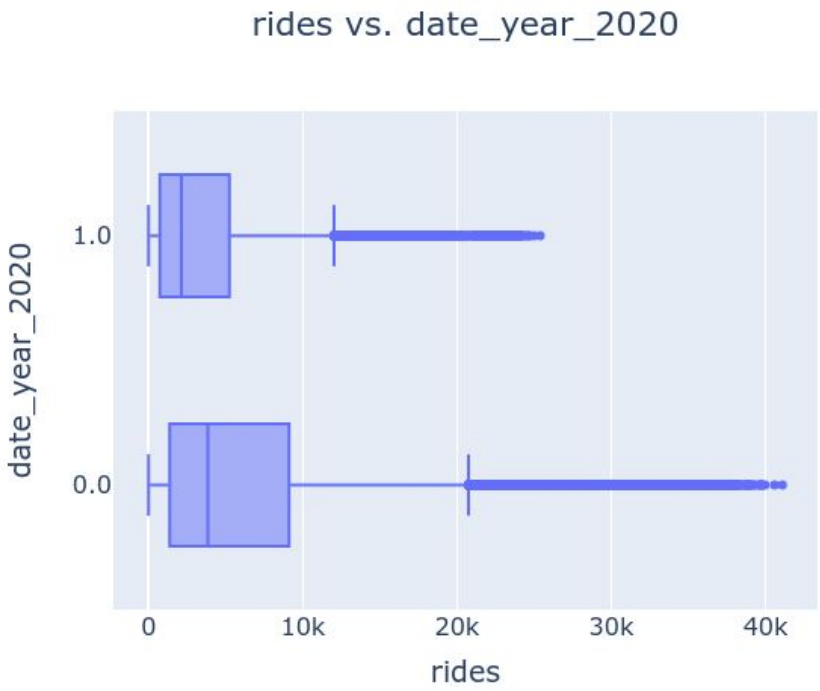
On average, the predictions are off by 820 to 880 riders per day, which is a tight range.



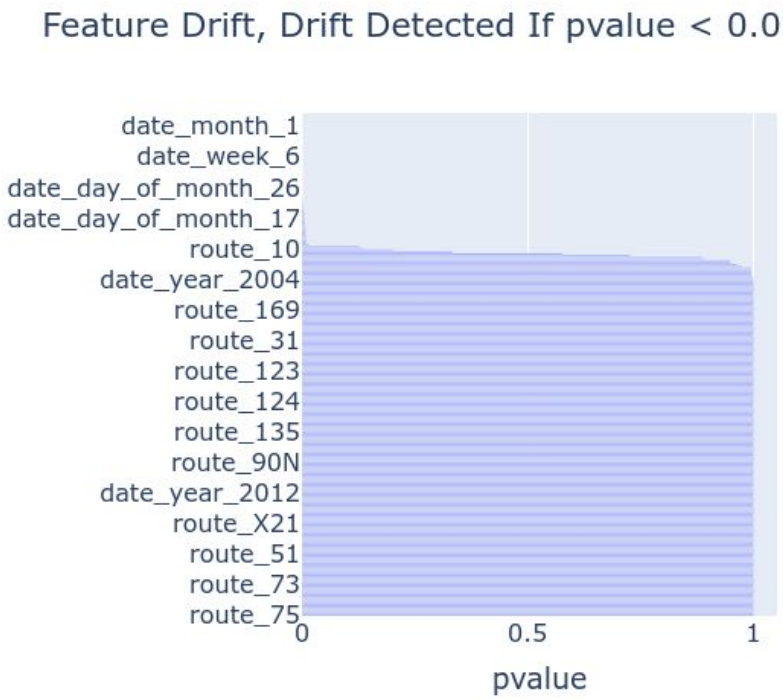
These are the top ten most important indicators of rider volume. The year 2020, the previous day of rider volume, route 79, and GDP show significantly more importance, and then the importance trails off.



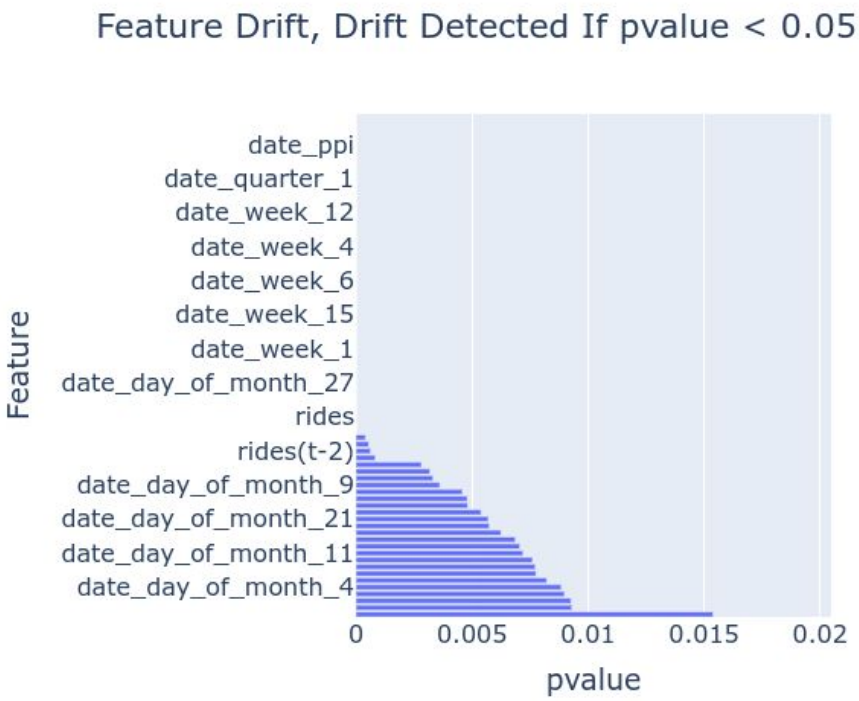
We can see a significant difference in rider volume where route 79 has a median of 25.8 thousand riders, and other routes have a median of 3687 riders.



We can see the year 2020 had less riders, likely due to COVID. The median for 2020 was 2122 riders, and the median for other years was 3838.

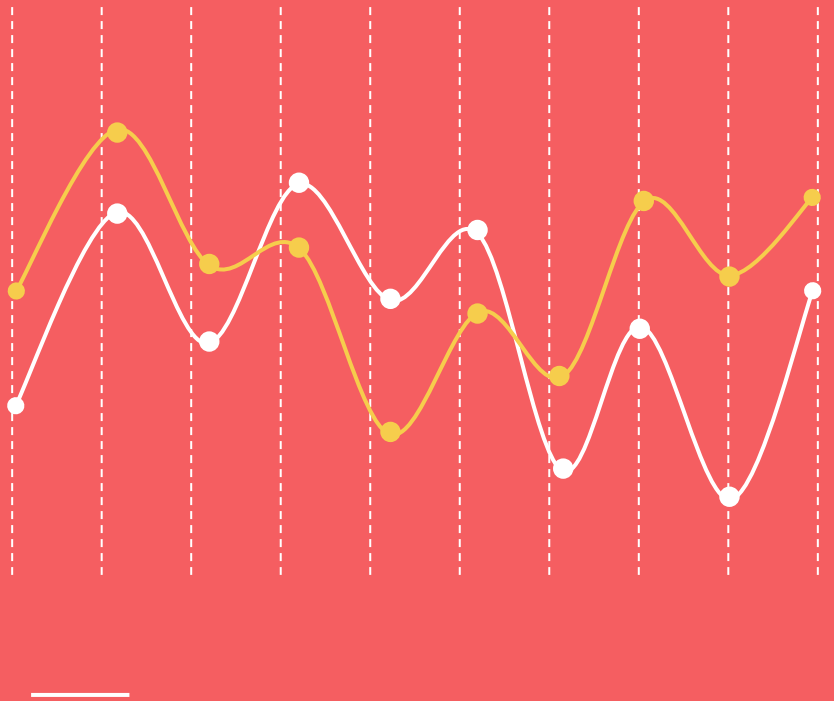


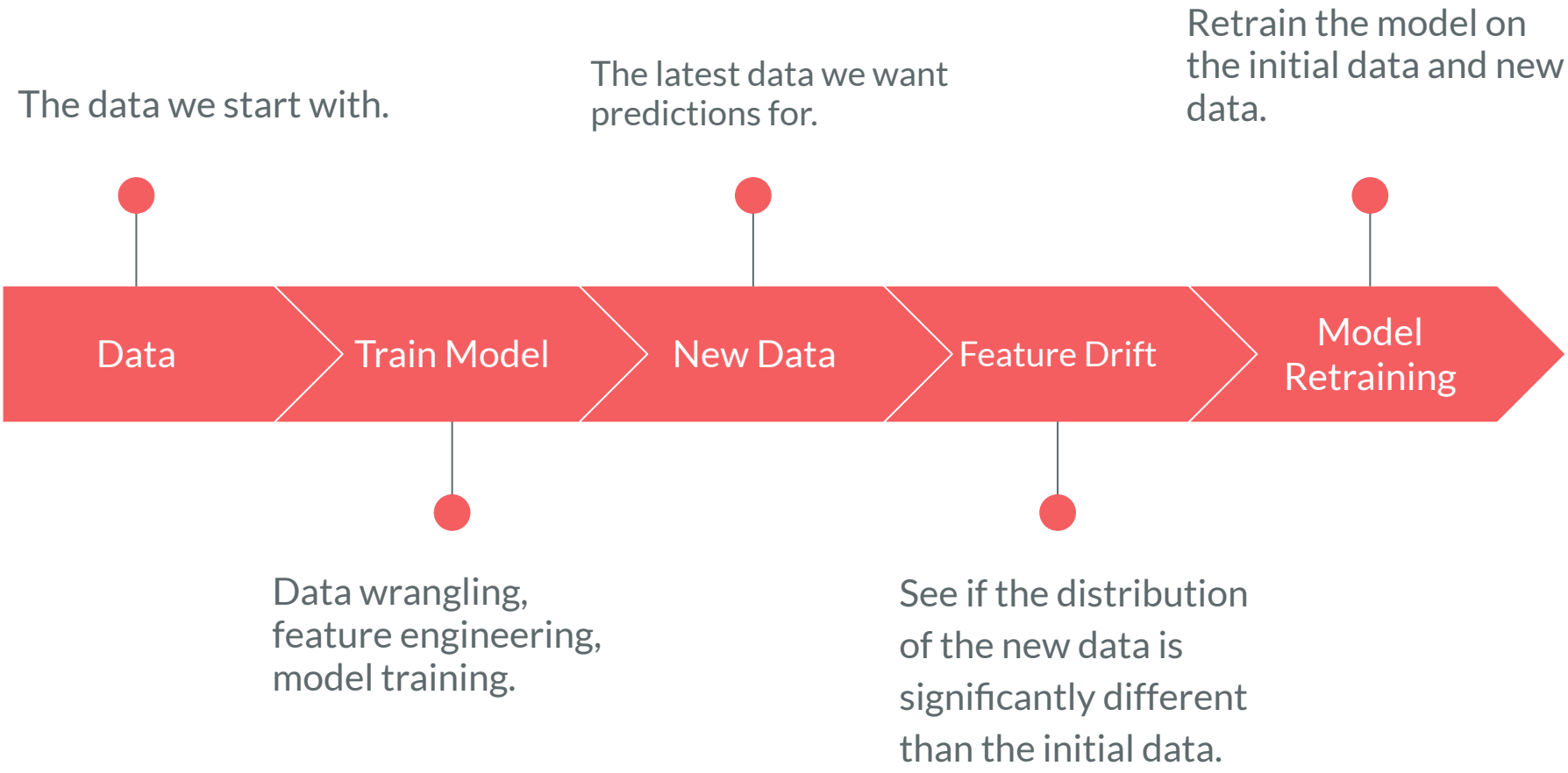
A Kolmogorov-Smirnov test was performed for each column in the data to see if the distribution of the testing data is the same as the training data. If the testing data does not share the same distribution as the training data, then there is a drift, which signals for model retraining. Most of the columns do not experience a drift, which is good.



These are the features which are experiencing a drift. They consist of timestamp components, rider volume, and economic indicators. This is because time is moving forward, rider volume is changing over time, and economic indicators have a trend.

Deployment




The data we start with.
The latest data we want predictions for.
Retrain the model on the initial data and new data.
Data
Train Model
New Data
Feature Drift
Model Retraining
Data wrangling, feature engineering, model training.
See if the distribution of the new data is significantly different than the initial data.

Thank You