Chicago Subway ("L") Ridership: Comparing Forecasting Methods

- Introduction
- Feature Engineering
- Forecast Overview
- Results
- Future Plans
Introduction
L Data

API accessed with RSocrata library

Ridership (by station by day)
Station Info (e.g., lat & lon)
Divvy Data

Programmatically downloaded .csv & .xlsx files from the Divvy website (https://www.divvybikes.com/system-data)

Trip Details (start location, stop location, datetime start, datetime stop, user type, etc)

Station Info (e.g., lat & lon, station "in use" date, etc.)
Holiday Data

Data scraped from https://www.officeholidays.com with rvest library

Holiday Date, Holiday Name, Comment, etc.
Weather Data

Request made on https://www.ncdc.noaa.gov/cdo-web/ and was emailed .csv files

Date, temperature max, precipitation, snow depth, etc.
Objectives

1 Week Ahead L Forecasts (by station by day)

Compare Forecasts By Algorithm

Explore Variable Effects
Chicago Subway ("L") Ridership: Comparing Forecasting Methods

Introduction

Feature Engineering

Forecast Overview

Results

Future Plans
Feature Engineering

44 predictor variables (131 after one-hot encoding)
Distance

# of Divvy stations with 0.5 miles of an L station

miles from an L station to the closest Divvy station
Divvy Trips
(done daily for each of the three types of customers)

- trip counts
- trip time stats (e.g., mean, median)
Holidays

holiday name and date
L Ridership

lags of ridership

moving averages of ridership
Time

day of the week

month

week of the month
Temperature

minimum daily temperature (in 25 F bands)
maximum daily temperature (in 25 F bands)
Chicago Subway ("L") Ridership: Comparing Forecasting Methods

- Feature Engineering
- Results
- Forecast Overview
- Future Plans
- Introduction
Forecast Overview

Sliding Window

Expanding Window

Comparison Criteria

Models

Procedure
Procedure

caret, rsample, purrr

- Model for each station (143/6)
- Train/Valid/Test (56/24/20)
- Training uses time-slice validation (1.5/0.5/13)

- One-hot encoding
- Near zero variance
- High correlation
- Centering
- Scaling
# Models

<table>
<thead>
<tr>
<th>library::model</th>
<th>variables used</th>
</tr>
</thead>
<tbody>
<tr>
<td>caret::randomForest::randomForest</td>
<td>All (after OH, &amp; NZV)</td>
</tr>
<tr>
<td>caret::xgboost::xgboost</td>
<td>All (after OH, NZV, &amp; HC)</td>
</tr>
<tr>
<td>forecast::auto.arima</td>
<td>date, rides</td>
</tr>
<tr>
<td>forecast::auto.arima</td>
<td>date, rides, fourier transformation, external regressors identified by RF &amp; XGBTree</td>
</tr>
<tr>
<td>prophet::prophet</td>
<td>date, rides</td>
</tr>
<tr>
<td>prophet::prophet</td>
<td>date, rides, holidays</td>
</tr>
<tr>
<td>h2o::h2o.automl</td>
<td>All (after OH)</td>
</tr>
<tr>
<td>h2o::h2o.automl</td>
<td>All (after OH, NZV, &amp; HC)</td>
</tr>
</tbody>
</table>
Comparison Criteria

Accuracy (RMSE) on validation data

Run time
Chicago Subway ("L") Ridership: Comparing Forecasting Methods

- Feature Engineering
- Results
- Introduction
- Forecast Overview
- Future Plans
Results
Extrap vs. Interp

- **Extrapolation**
- **Interpolation**

RMSE vs. start_date for different model types:
- **ARIMA**
- **h2o**
- **prophet**
- **rf**
- **arima_xreg**
- **h2o_limearse**
- **prophet_hol**
- **xgb**
Champion?
Chicago Subway ("L") Ridership: Comparing Forecasting Methods

- Feature Engineering
- Forecast Overview
- Results
- Future Plans

Introduction
Future Plans

- Keras (LSTM)
- Move computation to an AWS GPU image
- Expanding-window time-slice
- Accuracy as MASE
- Forecast & effects on Divvy ridership
Contact Info

turse.mda@gmail.com
https://www.linkedin.com/in/mdaturse/
https://github.com/supermdot/Chicago_EL_Divvy
http://cubs.com/mdaU/
Chicago Subway ("L") Ridership: Comparing Forecasting Methods

- Introduction
- Feature Engineering
- Forecast Overview
- Results
- Future Plans