Data Scientist Challenge

Data Cleaning

The dataset was cleaned from outliers to remove abnormal conditions. The minimum and maximum values used to remove outliers are summarized in the table below.

Feature	Min	Max
Temperature	-20	
Windchill_Index	-30	
Wind_Speed		50
Humidex		45
Dew_Point	-22	

Preprocessing

The dataset was preprocessed through feature engineering. Datetime features such as day of the week, month number and day were created using the date column. Another feature `Holiday` was created to indicate whether the date was a holiday. The `Ontario_Demand` target tag was shifted back 24 hours to setup the dataset to use past data to predict future demand. Lastly, the dataset was split into training and test sets using the date July 1 2020.

Model selection

Initial linear, tree and support vector regression models were trained with default parameters on the training dataset. The resulting mean absolute percentage error is shown in the table below.

Model	MAPE
ExtraTreesRegressor	0.0453226
HistGradientBoostingRegressor	0.0454092

Model	MAPE
RandomForestRegressor	0.046697
GradientBoostingRegressor	0.0509691
DecisionTreeRegressor	0.0615749
Ridge	0.0865246
Lasso	0.0866869
SVR	0.103891
ElasticNet	0.106966

The linear and support vector regression based models had the worse performance compared to the tree based models. Thus, tree based models were selected for further fine tuning.

Hyperparameter tuning process

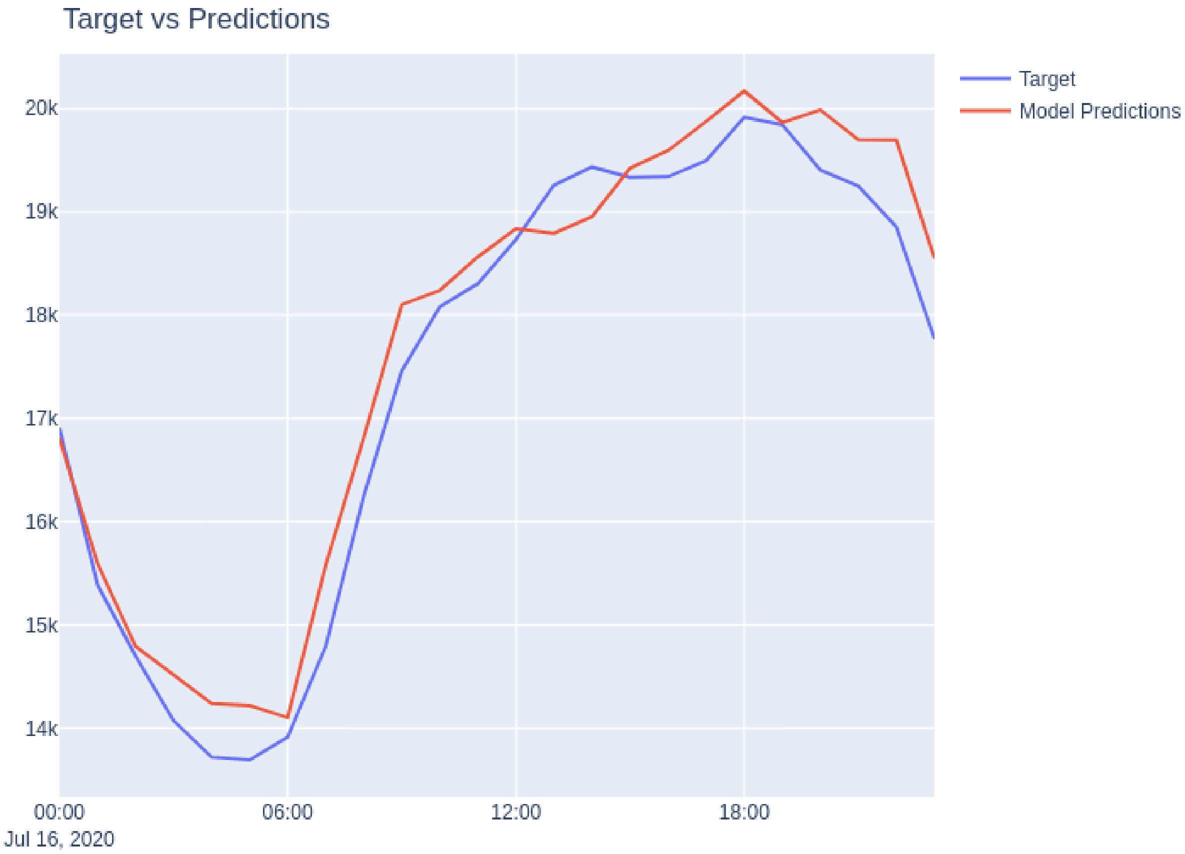
The histgradientboosting model was selected for further hyperparameter tuning. The hyperparameter search space is shown below. Randint is a random integer distribution ranging from a low value to a high value. Uniform indicates a uniform distribution ranging from "loc" value to "loc" + "scale" value.

```
{"learning_rate":uniform(loc=0.0001,scale=1),
    "max_leaf_nodes":randint(low=1,high=20),
    "max_iter":randint(low=500,high=1500),
    "min_samples_leaf": randint(low=5,high=30),
    "max_bins":randint(low=3,high=255),
    "l2_regularization": uniform(loc=0,scale=2)
}
```

Random Seach with Cross validation was used to tune the models using the search space. A time series cross validation split of 5 periods was used to split the dataset and calculate metrics. The random search was conducted for 100 iterations. After hyperparameter tuning, the test set MAPE value was 0.044092.

Results

The target and predicted model results are shown for June 16 2020. The model predicted each value using the data from 24 hours ago.


Target vs Predictions
Target
Model Predictions
20k
19k
18k
17k
16k
15k
14k
00:00
Jul 16, 2020
06:00
12:00
18:00