Avocado Case Study

Forecasting avocado prices and predicting sales

Problem: predict avocado prices and sales.

The data:

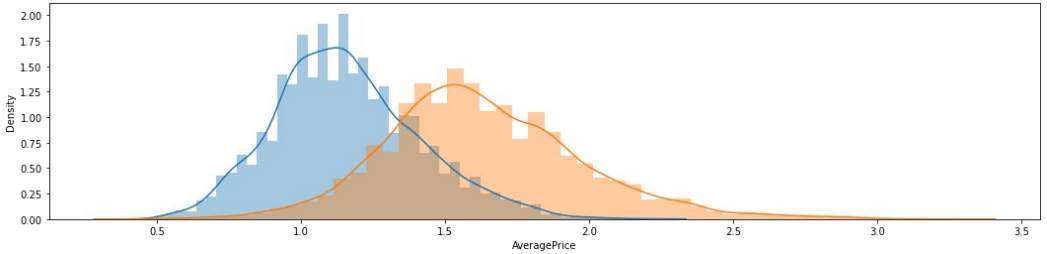
- 25000+ rows of prices and sales/type/region/week.
- Google search data for avocado related searches for the US- total number of searches/week.
- Data over 233 weeks from 4.01.2015 to 14.07.2019
- 54 total regions, each with 466 entries.

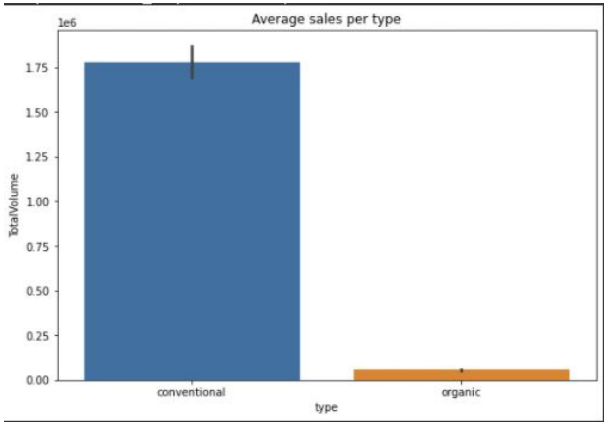
	Date	AveragePrice	TotalVolume	4046	4225	4770	TotalBags	SmallBags	LargeBags	XLargeBags	type	year	region	avocado	organic	recipe	toast	sandwich	organic_avocado
0	2015-01-04	1.22	40873.28	2819.50	28287.42	49.90	9716.46	9186.93	529.53	0.00	conventional	2015	Albany	46.0	76.0	84.0	8.0	51.0	10.0
1	2015-01-04	1.00	435021.49	364302.39	23821.16	82.15	46815.79	16707.15	30108.64	0.00	conventional	2015	Atlanta	46.0	76.0	84.0	8.0	51.0	10.0
2	2015-01-04	NaN	788025.06	53987.31	552906.04	39995.03	141136.68	137146.07	3990.61	0.00	conventional	2015	BaltimoreWashington	46.0	76.0	84.0	8.0	51.0	10.0
3	2015-01-04	1.01	80034.32	44562.12	24964.23	2752.35	7755.62	6064.30	1691.32	0.00	conventional	2015	Boise	46.0	76.0	84.0	8.0	51.0	10.0
4	2015-01-04	1.02	491738.00	7193.87	396752.18	128.82	87663.13	87406.84	256.29	0.00	conventional	2015	Boston	46.0	76.0	84.0	8.0	51.0	10.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
25156	2019-07-14	1.66	4007.93	218.47	252.29	0.00	3537.17	1460.65	2076.52	0.00	organic	2019	Syracuse	60.0	76.0	68.0	53.0	50.0	43.0
25157	2019-07-14	1.06	3767.89	129.01	0.00	0.00	3638.88	3635.55	3.33	0.00	organic	2019	Tampa	60.0	76.0	68.0	53.0	50.0	43.0
25158	2019-07-14	1.99	1236969.18	106370.49	209820.63	5606.10	915082.38	667494.94	247562.25	25.19	organic	2019	TotalUS	60.0	76.0	68.0	53.0	50.0	43.0
25159	2019-07-14	2.33	209408.22	23918.57	42432.02	985.67	142071.96	75883.13	66163.64	25.19	organic	2019	West	60.0	76.0	68.0	53.0	50.0	43.0
25160	2019-07-14	1.84	16372.24	1195.71	681.10	2765.31	11730.12	10861.33	868.79	0.00	organic	2019	WestTexNewMexico	60.0	76.0	68.0	53.0	50.0	43.0

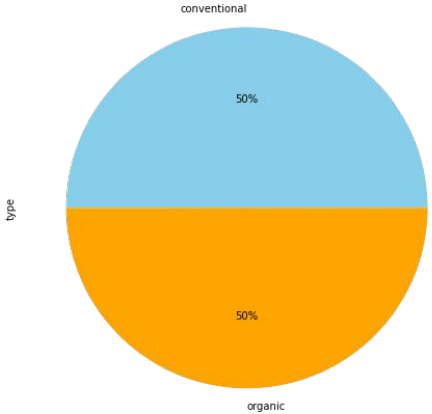
25161 rows × 19 columns

Key points in the data

- Average price distribution is higher for organic avocados
- The dataset contains equal entries for organic and conventional
- Average volume of sales is considerably lower for organic



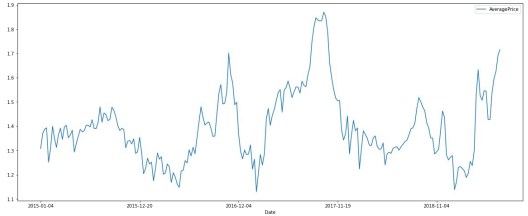


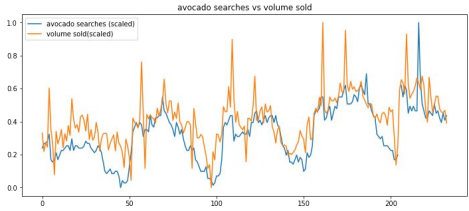


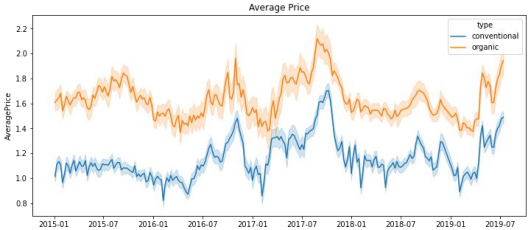
Time series exploration

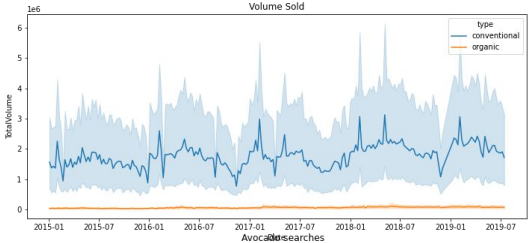
Key points in the data

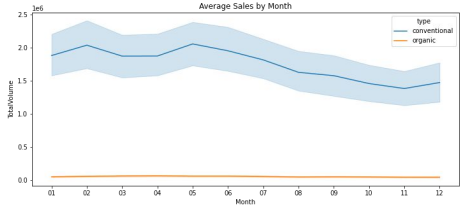
- Average price over time correlates with demand.
- The google searches containing the word “avocado” correlate with sales
- Avocado toast popularity is increasing over time
- There are monthly/season trends

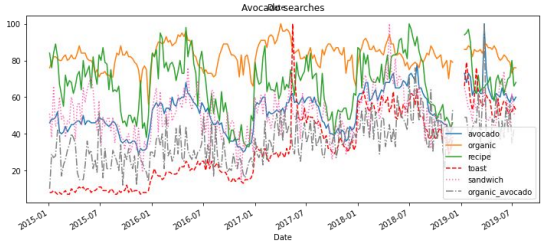


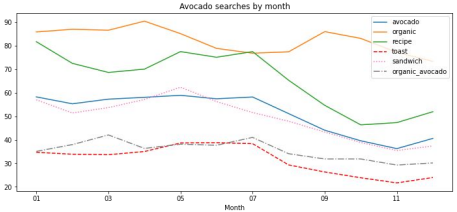








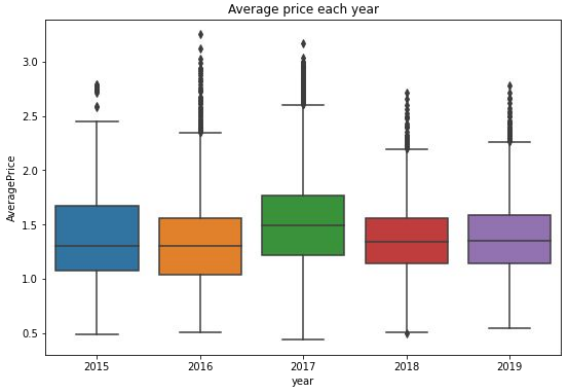


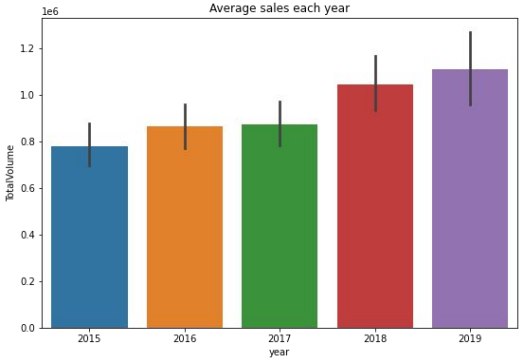


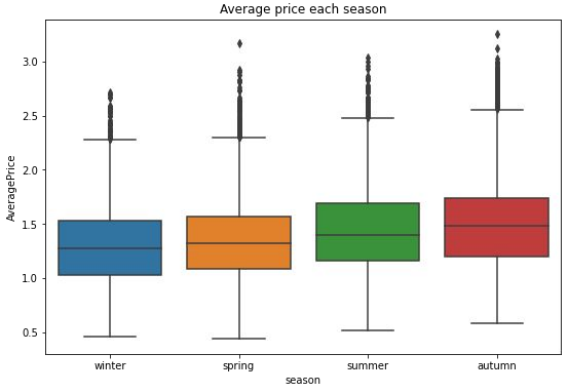
Seasonal exploration

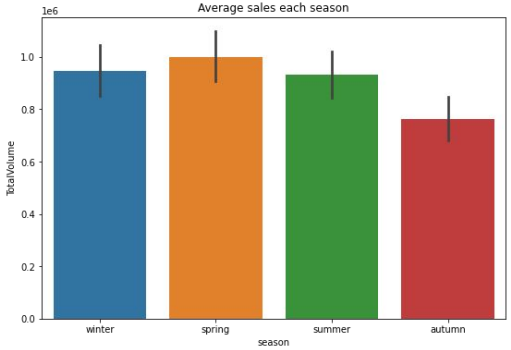
Key points in the data

- Introduced new variable: season
- Prices are highest in the autumn
- Sales are going up each year
- Sales are highest in the spring

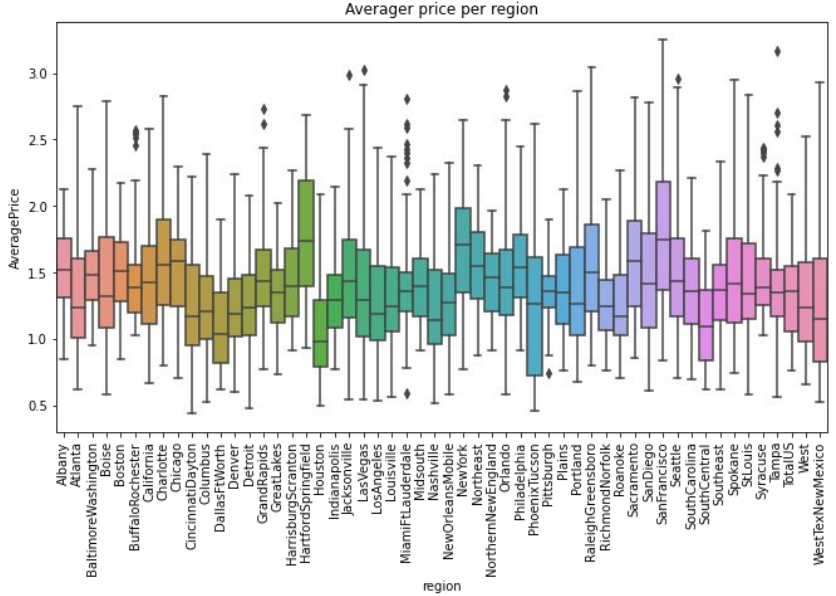


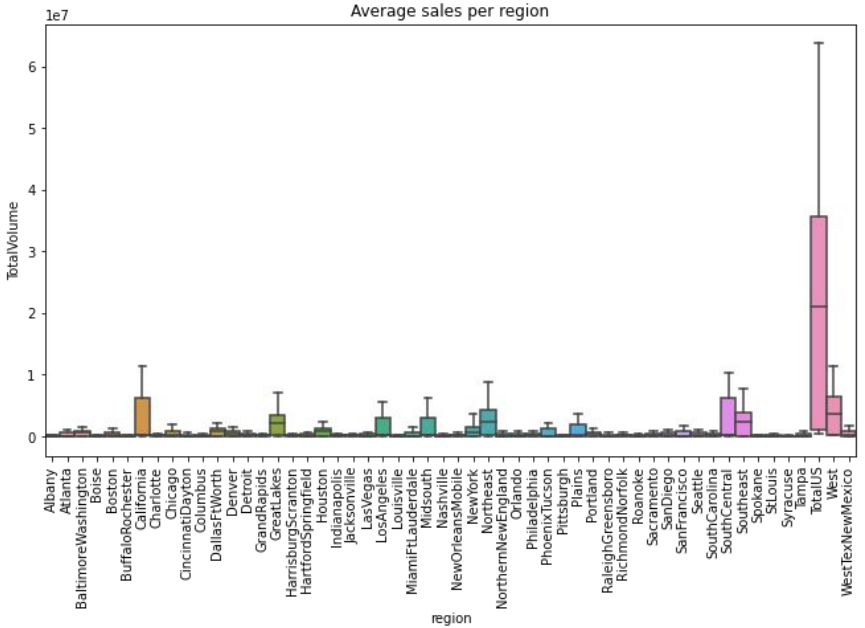






Regional exploration





Forecast

Forecasting avocado prices using historical price data

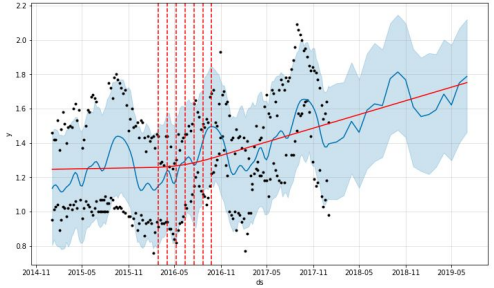
Facebook Prophet accounts for yearly seasonality.

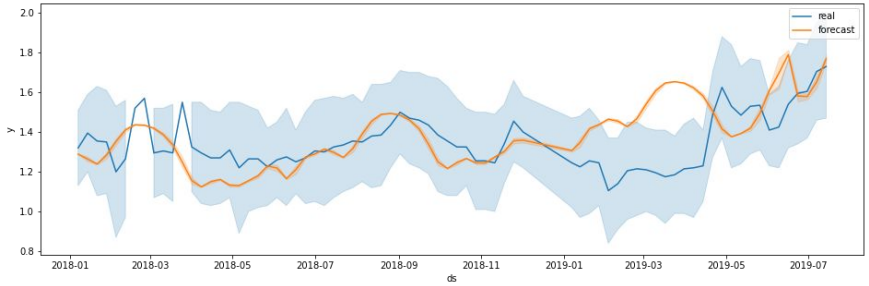
Avocado prices for 2018 and 2019, for the total US, have been forecasted using the price data from <2018.

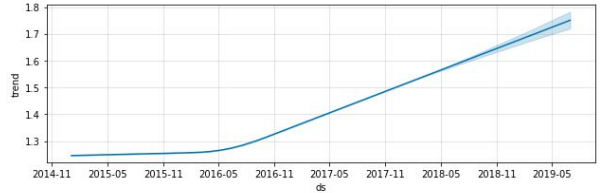
FB Prophet forecast

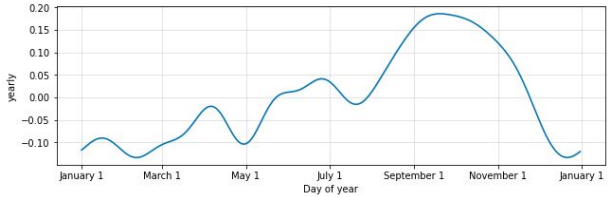
Method

- 466 price entries for totalUS from 2014 to 2019
- Training: 314 samples from 2014 - 2018.
- Testing: holdout set of 152 samples from 2018 and 2019
- Seasonality has been captured well.









Sales prediction

Predicting weekly avocado sales using google search data from the previous week.

Linear and non-linear models have been evaluated and the best model (Random Forest) has been chosen.

Sales prediction for “next week”.

Method

- Continuous Predictors: Average price and google searches from previous week
- Categorical predictors: type, region, month, season, year
- Response: ShiftedSales - volume of sales for the type and region, shifted by one week in advance.
- Data has been split into training (80%) and testing (20%)

Preprocessing

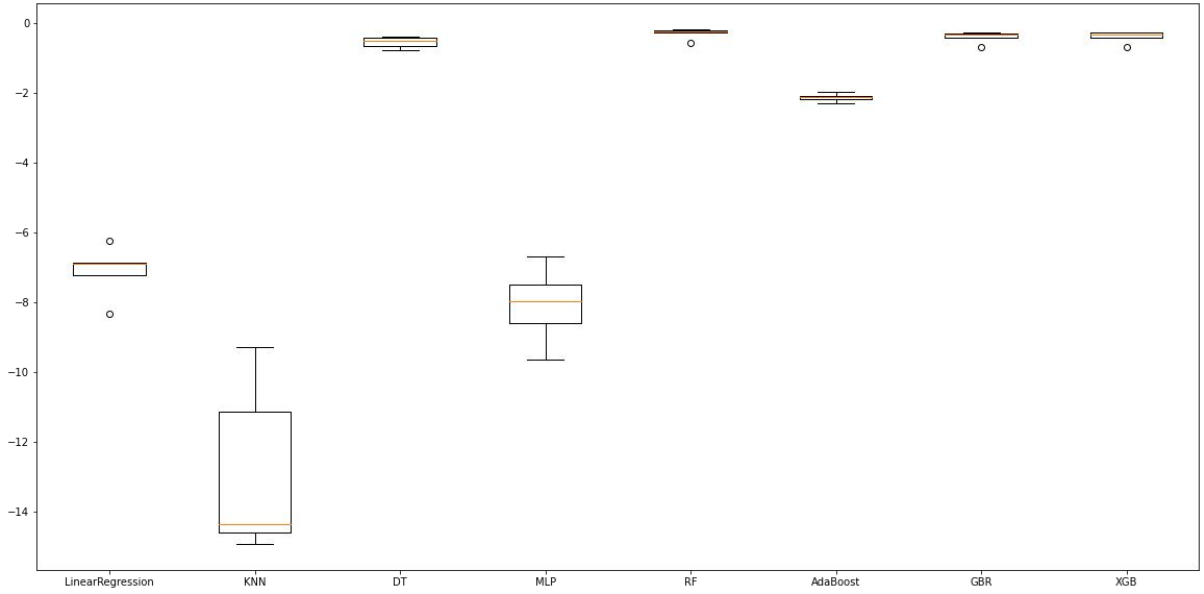
- Numeric variables have been scaled for mean 0 and sd 1
- Categorical variables have been converted to one-hot vectors
- After preprocessing we have 76 predictors.

```
['AveragePrice', 'type', 'avocado', 'organic', 'recipe', 'toast',
       'sandwich', 'organic_avocado', 'year', 'region_Atlanta',
       'region_BaltimoreWashington', 'region_Boise',
'region_Boston',
       'region_BuffaloRochester', 'region_California',
'region_Charlotte',
       'region_Chicago', 'region_CincinnatiDayton',
'region_Columbus',
       'region_DallasFtWorth', 'region_Denver', 'region_Detroit',
       'region_GrandRapids', 'region_GreatLakes',
'region_HarrisburgScranton',
       'region_HartfordSpringfield', 'region_Houston',
'region_Indianapolis',
       'region_Jacksonville', 'region_LasVegas',
'region_LosAngeles',
       'region_Louisville', 'region_MiamiFtLauderdale',
'region_Midsouth',
       'region_Nashville', 'region_NewOrleansMobile',
'region_NewYork',
       'region_Northeast', 'region_NorthernNewEngland',
'region_Orlando',
       'region_Philadelphia', 'region_PhoenixTucson',
'region_Pittsburgh',
       'region_Plains', 'region_Portland',
'region_RaleighGreensboro',
       'region_RichmondNorfolk', 'region_Roanoke',
'region_Sacramento',
       'region_SanDiego', 'region_SanFrancisco', 'region_Seattle',
       'region_SouthCarolina', 'region_SouthCentral',
'region_Southeast',
       'region_Spokane', 'region_StLouis', 'region_Syracuse',
'region_Tampa',
       'region_TotalUS', 'region_West',
'region_WestTexNewMexico', 'Month_02',
       'Month_03', 'Month_04', 'Month_05', 'Month_06', 'Month_07',
'Month_08',
       'Month_09', 'Month_10', 'Month_11', 'Month_12',
'season_spring',
       'season_summer', 'season_winter']
```

Choosing the best model

Method

- Models tried: Linear Regression, KNN, Decision Tree, MLP, Random Forest, AdaBoost, Gradient Boosted regression, Extreme Gradient Boosted regression
- 5-fold cross validation using negative MSE as performance metric
- The best model (lowest MSE) was RF with a MSE of 0.29



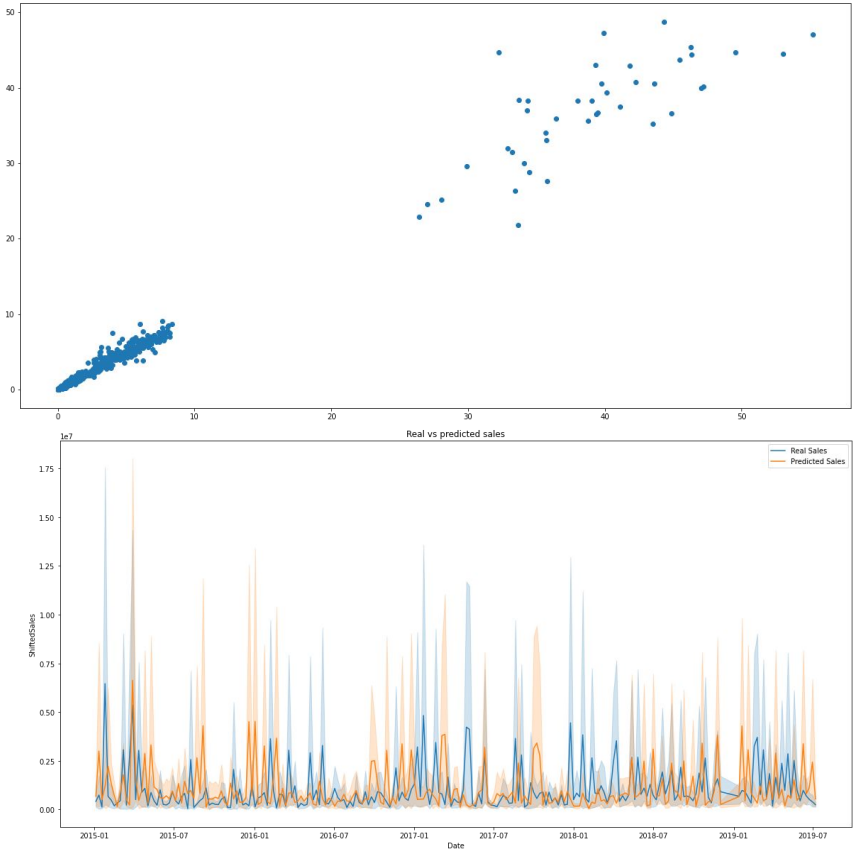
Testing the model

Method

- Model has not been tuned due to lack of time.
- Some turning options: RandomSearch, GridSearch, Bayesian optimization with CV
- Model Scores on the final test set, with the default hyperparameters and 500 estimators:

```
MSE model: 0.3533520403118462
```

```
RMSE model: 0.594434218658251
```



Conclusion

Positives:

- The models are able to capture the trends and direction of the prices and sales.
- The price forecast shows and increasing trend in the average avocado prices in the US
- Google searches, price and regions are good predictors for avocado sales.

Improvements:

- Outliers and missing values treatment
- More forecasting experiments with various other models
- Hyperparameter tuning
- Season, Month and year of “next week” could have improved results
- Day of the month as a predictor