

# **2020 DubStech Datathon**

## **Grocery Retail Store Prompt**

**Team UWO**

**2020-05-17**

## Setup

To start off with this problem, we first defined the metric we wanted to predict. In our case, it is known as cumulative sales for the next 91 days. We segmented the data on CATEGORY of product sold. This gave us all the receipts for a given group of goods (such as markdown bags). We then summed all the receipts on a given day, giving us a day by day look at the revenues a category of product brings into the store.

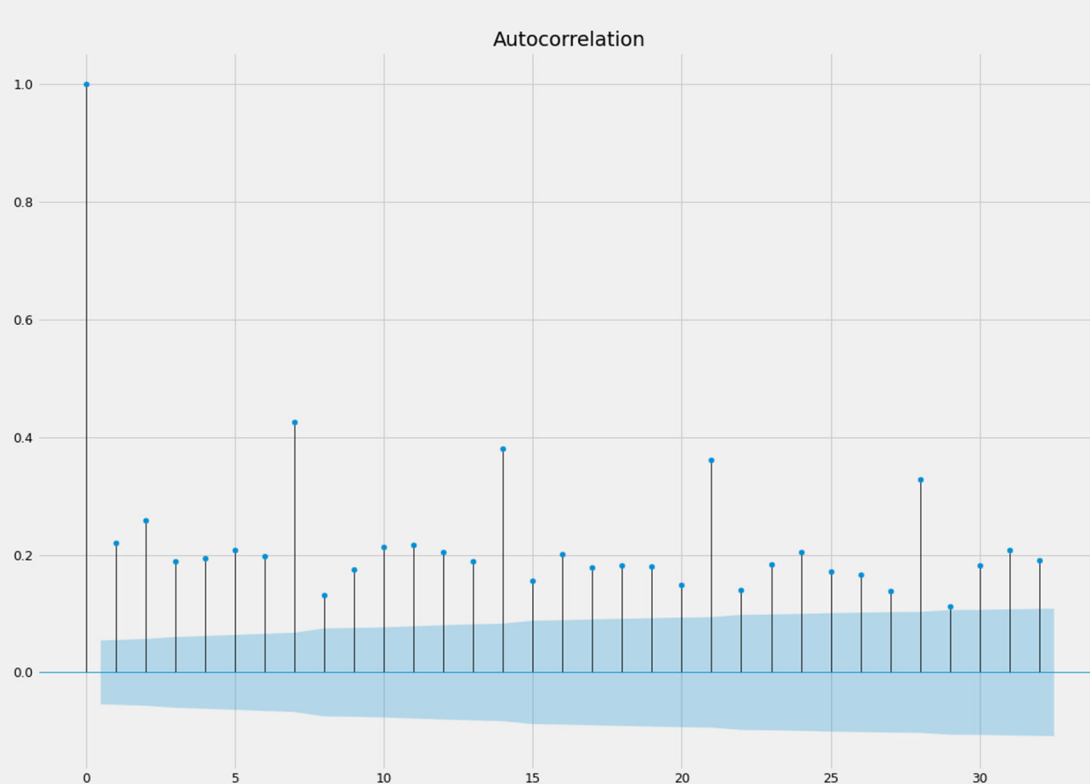
## Time Series:

To determine the validity of using a time series we first performed STL decomposition on TOTAL\_PRICESELL for all categories as well as the store wide revenue observed at a daily frequency. This led us to conclude that the time series did exhibit seasonality in quarters, and it is something we would have to account for when using a classical time series model.

Store Wide Revenue



To accompany our STL decomposition we also plotted the Autocorrelation function of the total store revenue time series to observe and pick which lags will have the most predictive power. As we can see lags 7, 14, 21 and multiples of 7 have predictive power. This agrees with anecdotal evidence that people buy groceries on a weekly time frame.



This led us to test the time series of TOTAL\_PRICESELL, adjusted for a 1-day frequency, for stationarity. We opted to use the SARIMA model, as a benchmark to other models as this model considers seasonality and had the lowest AIC out of other models compared such as ARIMA, ARMA.

### ***Feature Engineering:***

In order to build robust models for time series prediction, we utilized multiple feature engineering techniques to extract most information from our given dataset.

- Basic features:
  - Three features that will be used are TOTAL\_PRICEBUY, TOTAL\_PRICESELL, and PROFIT for each day.
- Time series features:
  - According to our autocorrelation tests, we noticed that past data do have predictability into the future. We computed rolling sum, mean, and standard deviation of the three basic features listed above, with different lookback periods such as 7, 14, 21 and 91 (approximately 3 months) days.
- Seasonality features:
  - According to our seasonality analysis, we noticed it plays a significant effect in our data. To allow regression models to take into account seasonality, we incorporated month and day-of-the-month as our features.

### ***Regression Models:***

We utilized various linear, ensemble, and tree models to determine the one that best predicts future sales. The models included in our research are listed below:

- Linear Regression
- Ridge
- SGD Regression
- Elastic Net
- Lars Regression
- Lasso
- Bayesian Ridge Regression
- Decision Tree Regression
- Extra Tree Regression
- Bagging Regression
- Random Forest Regression

### ***Feature Selection:***

For feature selection for each of the models (where applicable) we used recursive feature elimination with cross validation (k=5) to provide us a feature ranking that would then be applied on the data frame to keep the most relevant features. Because recursive feature elimination is a model-dependent algorithm we are able to extract the optimal set of features for each model.

### ***Model Selection:***

This set up our data to be inputting into various regression models in order to determine the one that best predicts future sales for each category of items. Given the range of various models, we computed in sample and out of sample MSE, using the out of sample value as the performance metric on which we selected our model.

*Table 1: Pumpkin Category regression models and their in-sample and out-of-sample MSE.*

Through the process, we were able to determine the best fitting models for each category and therefore able to compute total predicted sales effectively.

### ***Results***

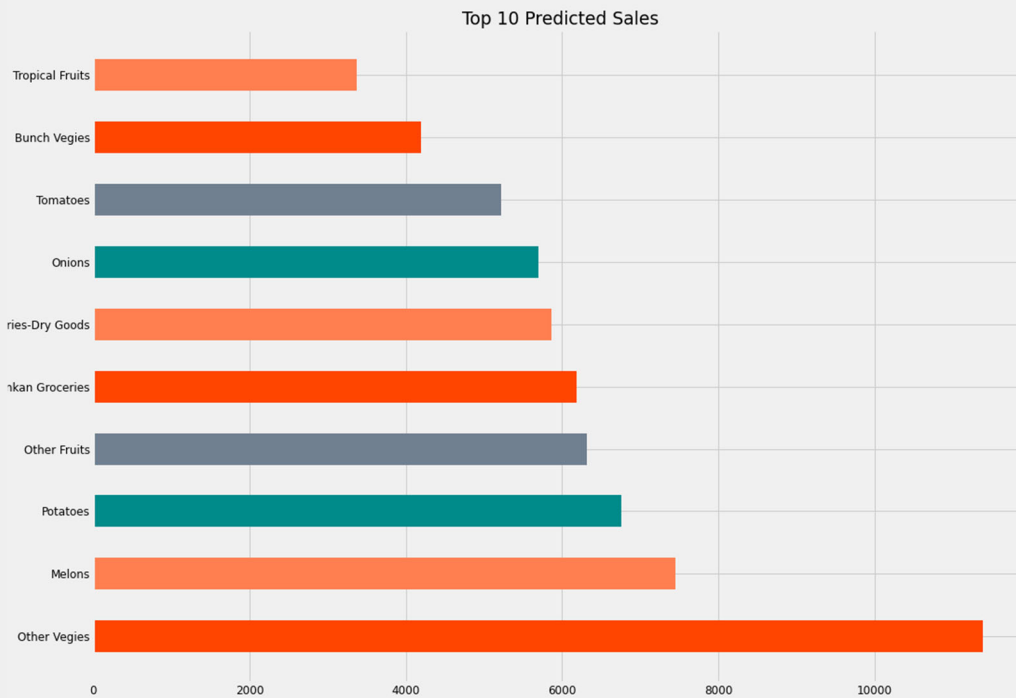
Attached below are the results from the analysis. We ranked the predicted sales for each category of items in order to infer the optimal decisions.

	<u>In-Sample MSE</u>	<u>Out-of-Sample MSE</u>	<u>Predicted 3 Month Sales</u>
Linear	199496.2967	96661.22637	2654.849868
Ridge	78647.85862	77330.49733	2509.80118
<b>SGD</b>	<b>1.48E+28</b>	<b>1.06E+28</b>	<b>-8.3153E+13</b>
ElasticNet	87291.40334	70370.63677	2492.926091
Lars	236336.699	147551.5407	2436.98829

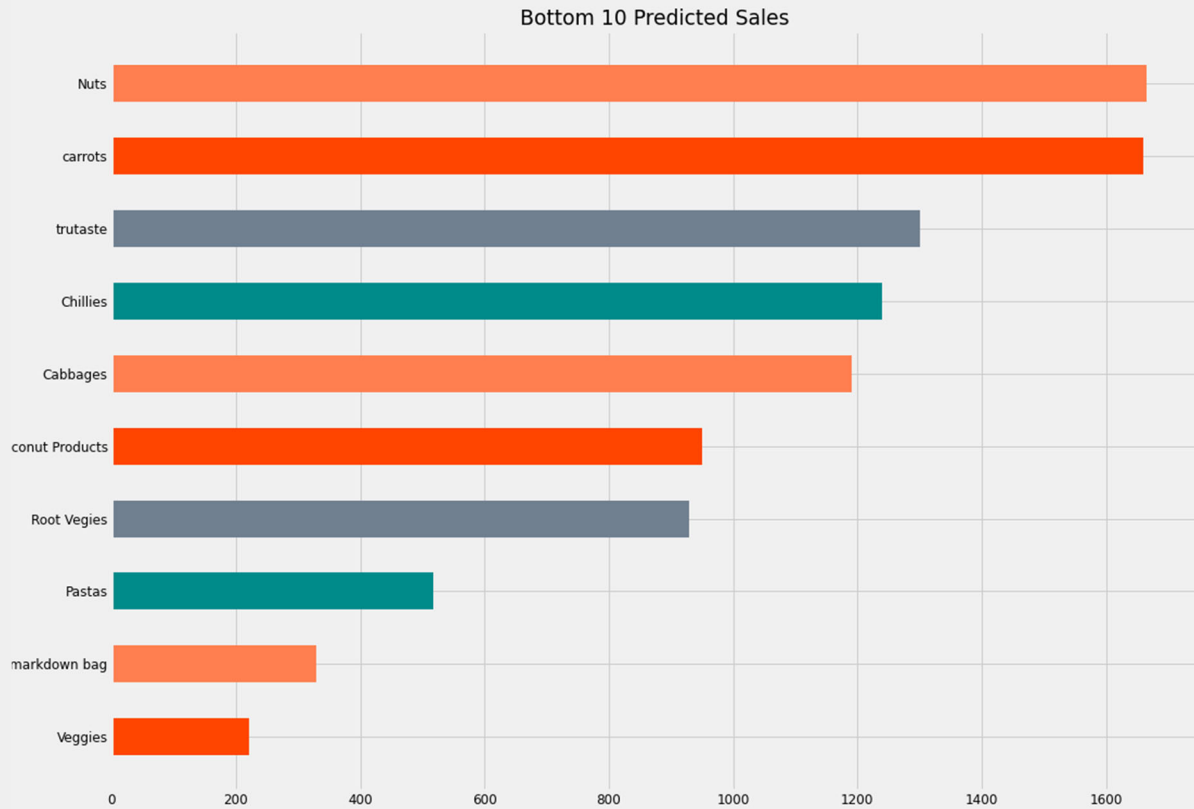
Lasso	82596.53612	78288.50642	2515.572223
BayesianRidge	98084.79061	92495.28315	2138.278966
DecisionTreeRegressor	0	150948.2229	2139.73755
ExtraTree	0	181784.6476	1787.00806
Bagging	1021.249631	131862.1406	2181.208951
RandomForest	849.3032922	161884.2526	2118.800163
SARIMA	25961.09145	989423.538	1539.845631

*Table 1: Pumpkin Category regression models and their in-sample and out-of-sample MSE.*

As we see from the plot below. *other veggies* ranked highest in terms of predicted sales, this should be ignored in our decision since it is just a bracket of miscellaneous items. Ranked second and third are *melons* and *potatoes*. Hence, it is recommended for the store to increase the inventory of these items in order to have enough supply to fulfill customer's demands.



As we see from the plot below. *Markdown bag*, *veggies* and *pastas* are expected to have the lowest predicted sales over the upcoming months. Hence it is recommended for the store to not allocate excess resources on these products in terms of both inventory and marketing expense.



To further support our analysis. We plotted the percentage of item sales for each category with 100% margin. Note that the reason behind 100% margin is that these items are grown on the farm of the grocery, so their acquisition cost of these items is \$0.00. We note the *potatoes* category, which was predicted to have the highest sales, also has a significant percentage of items with 100% margin. The *veggies* category, which was predicted to have the lowest sales, also consists of a low percentage of items with 100% margin.

