# CS677 - Final Project

Christopher Agostini

2023-12-14

## Summary

This project analyses sales data for a BBQ restaurant, builds a multivariate multiple regression model and is used to forcast the sales of categories of items in the future.

## Data Set

Historical data of sales is publicly available on the internet and includes items sold between May 01 to Dec 07 of 2023.

The data set can be downloaded from this URL: https://agsc-cs677-final-project.s3.us-east-2.amazonaws.com/items.csv

### Analysis

The following analysis and development was conducted on the original data:

* Determine which columns were useful as features, predictors and not required
* Design an intermediate data frame that combined items and modifiers
* Identify and categorize items by product type (with a targeted approach on meats)
* Organize original data set into 3 main cases and 7 special cases in comparison of the items, modifiers and prices
* Partition 28 subsets of data from the original data set, categorized by case
* Modularize Python code to enable sharing of all previously realized values as global variables throughout project
* Develop Python function to determine which case a line item belonged to and return item and price data
* Identify other features not part of the original data set that may influence the predictors

### Pre-Processing

The following processes were performed against the original data set in order to create an intermediate data frame:

* Drop columns form original data set that were not required
* Mutate price columns to correct float values and replacing null values
* Mutate the modifiers column by extracting text using regex and merging into the items and prices column
* Modify prices for modifiers depending on the type of items and expected portion size
* Generate datetime index and creating date, day, and time columns by parsing the original timestamp values

### Post-Processing

The final stage of generating the final data frame used for analysis required the following:

* converting day columns to binary variables and dropping other datetime columns besides index
* add additional feature columns as described in the features section below
* generate predictor columns from items and cost columns in intermediate data frame

### Features

The following features are used in the final data frame:

| Feature | Data Type | Description |
| --- | --- | --- |
| D01 | Binary | Day of week is Sunday |
| D02 | Binary | Day of week is Monday |
| D03 | Binary | Day of week is Tuesday |
| D04 | Binary | Day of week is Wednesday |
| D05 | Binary | Day of week is Thursday |
| D06 | Binary | Day of week is Friday |
| D07 | Binary | Day of week is Saturday |
| L01 | Float | Total sales lagged by 1 week |
| L02 | Float | Total sales lagged by 2 weeks |
| L03 | Float | Total sales lagged by 3 weeks |
| L04 | Float | Total sales lagged by 4 weeks |
| L05 | Float | Total sales lagged by 5 weeks |

### Predictors

The following predictors are prices of the categories of meat and other items sold.

| Predictor | Data Type | Description |
| --- | --- | --- |
| M01 | Float | Sales of brisket |
| M02 | Float | Sales of pork |
| M03 | Float | Sales of chicken |
| M04 | Float | Sales of ribs |
| M05 | Float | Sales of sausage |
| M06 | Float | Sales of turkey |
| M07 | Float | Sales of meat of unknown portion or type |
| O01 | Float | Sales of all other items not meat |

### Preview

A preivew of the final data set used in the machine learning is as follows:

img.png

## Machine Learning Model

Multi-Task Lasso regression from the Python Scikit-learn package is used to develop the machine learning model. 20% of the final data frame was selected randomly and used as training data. The model is then tested using the later 80% of data. The performance metric used to measure the accuracy of the model is the mean squared error and coefficient of determination (R^2).

### Results

| Product Category | Mean Squared Error | Coefficient of Determination |
| --- | --- | --- |
| Brisket | 166,249 | -0.239 |
| Pork | 14,586 | 0.5388 |
| Chicken | 3314 | -0.186 |
| Ribs | 12,291 | 0.5235 |
| Sausage | 1562 | 0.4588 |
| Turkey | 1531 | -0.433 |
| MeatOther | 492 | -0.2714 |
| Other | 113088 | 0.0525 |

### Conclusion

Future enhancements to this project will include adding additional features (e.g. weather forcast, expected volume, holidays, etc.) and conduct a more in depth review of which are best to use in the final model. Additionally, will use additional models to find which has the best performance metrics in addition to plots to spot outliers in the data and determine if transformations of data is required.

## Python

### Requirements

Before running this Python package and any of the child packages/modules, all packages listed in the requirements.txt should be installed into the Python interpreter environment used to run the code.

### Packages

The python code for this project is organized into multiple packages, with the top-level package only containing a main module which will execute all child packages and modules, and can be run by simply executing Python against a compressed version of this project.

#### Top-Level

The top-level package’s main module executes the child packages and modules in the order and manner required to download the original data set, perform the pre- and post-processing of the data to create the final data frame. Additionally, builds the models, performs the regression analysis, outputs the graphs and final results at the end.

### Modules

The child packages are developed into the programs components which further contain the code into respective modules. The first component is the configurations package, namely, the config package, which contains only the init module. Within this module, all global variables and parameters are saved and referenced here throughout every other package/module.

For detailed information about each additional package, please reference the top-level package’s main module for execution order. Likewise, see the doc string included in the modules for each additional package.