Supervised Learning Capstone – The Bread Basket

BY: CHRIS GARLAND

THINKFUL, FEBRUARY 2019

The Bread Basket Bakery

- Located in the historic city center of Edinburgh, Scotland
- Café offers specialty Argentine and Spanish baked products:
  - Medilunas
  - Empanadas
  - Alfajores

Menu

- Breakfasts
  - Coffee, tea, hot cocoa
  - Spanish Bakery Items
  - English Bakery Items
- Lunch
  - Sandwiches
- Seasonal
  - Christmas cakes (Pantone)
  - Bread Pudding

Research Question: Can we predict how many baked products need to be produced?

- Bakeries need provide fresh products daily
  - Keeps shop competitive in the marketplace (Grocery stores, other shops)
  - Customer expectations – higher quality than high shelf life products
- However, Bakeries need to control food waste/shortage
  - Having too much product:
    - Throwing away unsold product is a waste of money/time
  - Having too little product to sell:
    - Customers will be disappointed in the selection

Our Dataset – Strengths and Limitations

- The data is a log of each item sold from Oct 30th, 2016 around 10am through April 9th, 2017 around 3pm.
- Over 21,000+ items sold over this time frame.
- Each item is linked to a transaction number
- 95 unique items*
  - *Not really sure if each item is ACTUALLY unique but the cashier rang it up that way
  - Also some items listed as NONE
- Source: Kaggle

```
In [3]: # Data Exploration
        df.head()
```

Out[3]:

	Date	Time	Transaction	Item
0	2016-10-30	09:58:11	1	Bread
1	2016-10-30	10:05:34	2	Scandinavian
2	2016-10-30	10:05:34	2	Scandinavian
3	2016-10-30	10:07:57	3	Hot chocolate
4	2016-10-30	10:07:57	3	Jam

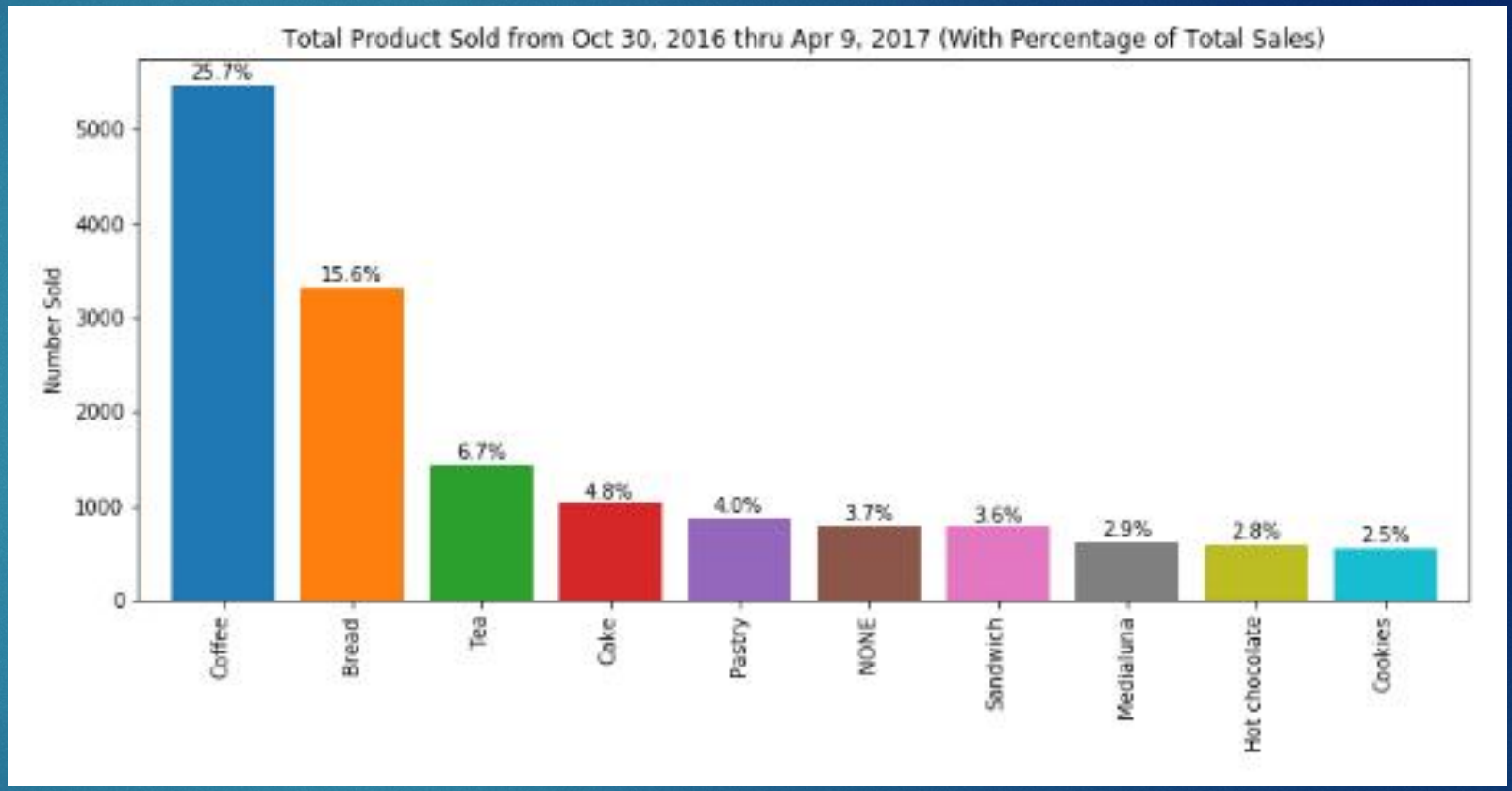
```
In [4]: df.tail()
```

Out[4]:

	Date	Time	Transaction	Item
21288	2017-04-09	14:32:58	9682	Coffee
21289	2017-04-09	14:32:58	9682	Tea
21290	2017-04-09	14:57:06	9683	Coffee
21291	2017-04-09	14:57:06	9683	Pastry
21292	2017-04-09	15:04:24	9684	Smoothies

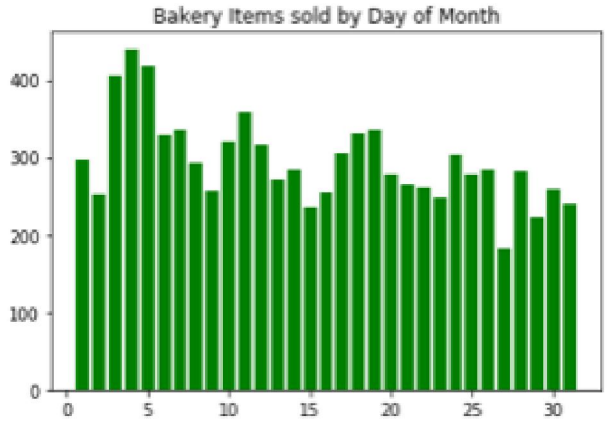
Top 10 Items Sold in the Data

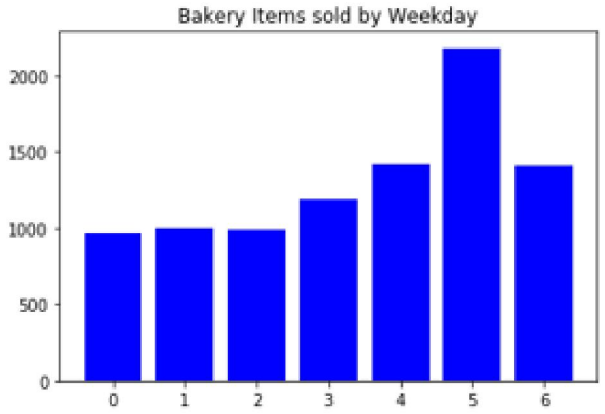
- Coffee is the top product sold and represents 25.7% of all items sold.
- Bread, Cake, Pastry and Medialuna, and Cookies are all represented in the Top 10.
- Less than 40% of items sold were a 'Baked Item' which is the focus of this project.

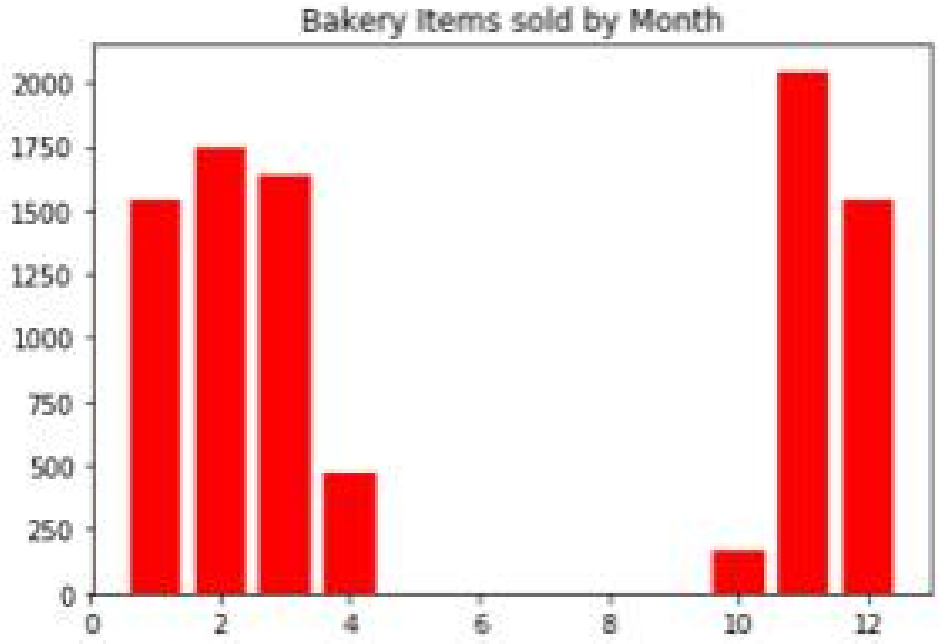


Focus on Bakery Items – Amount Sold by Date

- There were much more bakery items sold on Saturday than any other day of the week.
- Bakery items by month is similar. Months 4 and 10 had incomplete data.
- There were some peaks and valleys when looking at Day of Month but nothing really stood out.







Feature Engineering - Whats important for prediction?

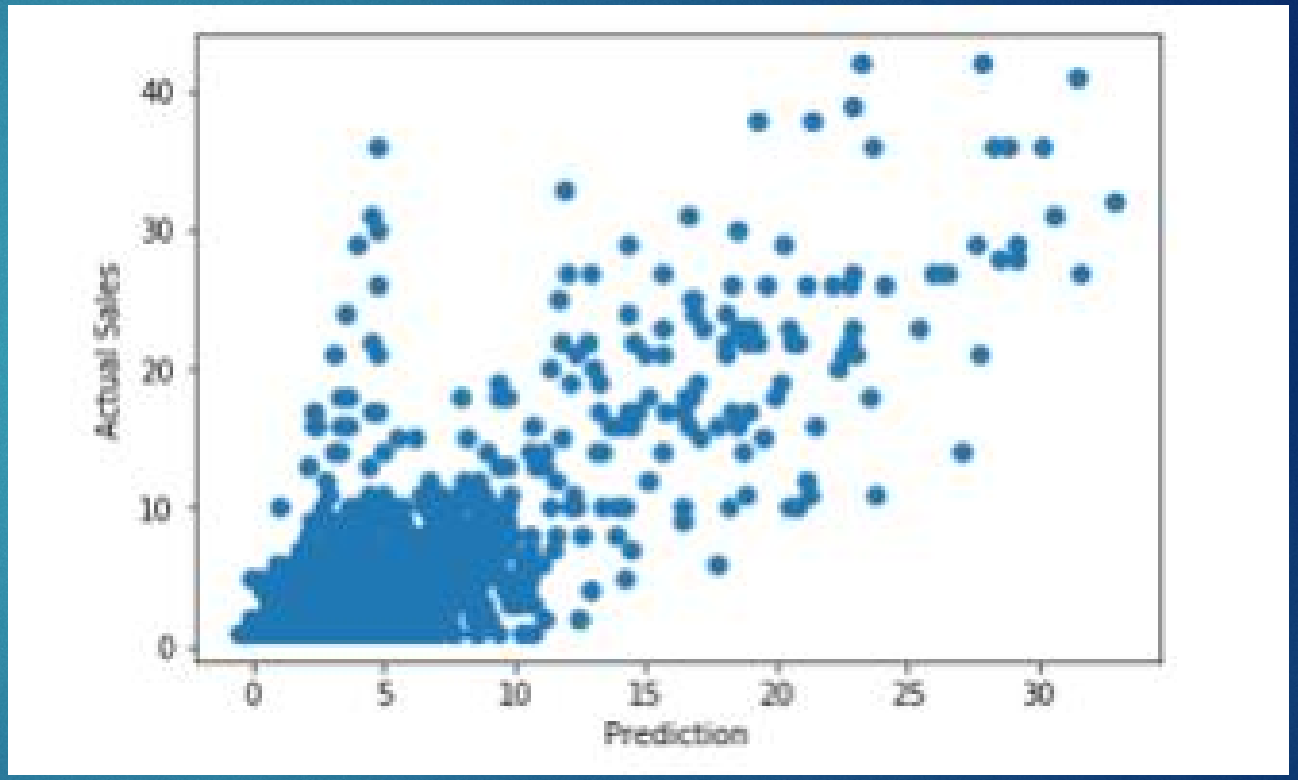
- Less than 40% of items sold were a 'Baked Item'
  - Need to isolate the bakery items for the model
- The date is important
  - Weekdays were especially important to predictability
- Previous weeks data
  - Fresh departments in grocery stores rely on last week sold and last year sold numbers to predict how much product to buy
  - Used 'Last Week' and '2-weeks prior' sales data as model inputs

Modeling - Regression

- Train/Test split
  - Trained data with date, items sold, last week items, 2 weeks ago items sold
  - Test data - Held out data from March to validate the model
- Used 3 different models to find the best performance
  - Simple Linear Regression
  - Random Forest Regressor
  - Gradient Boosting Regressor
- Optimized parameters with GridSearchCV
- Model Score and Mean Squared Error were used to judge model performance

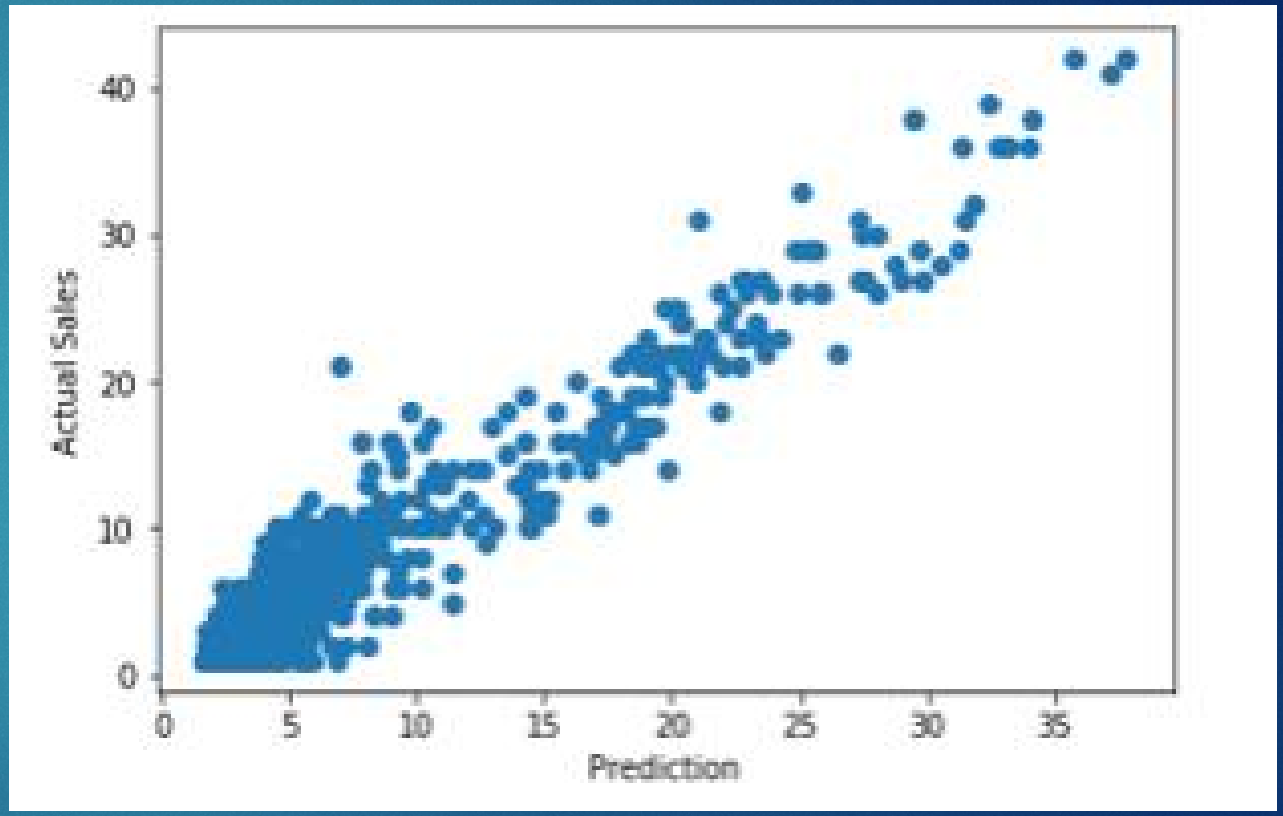
Simple Linear Regression Model

- Score 0.59
- Mean Squared Error 17.35



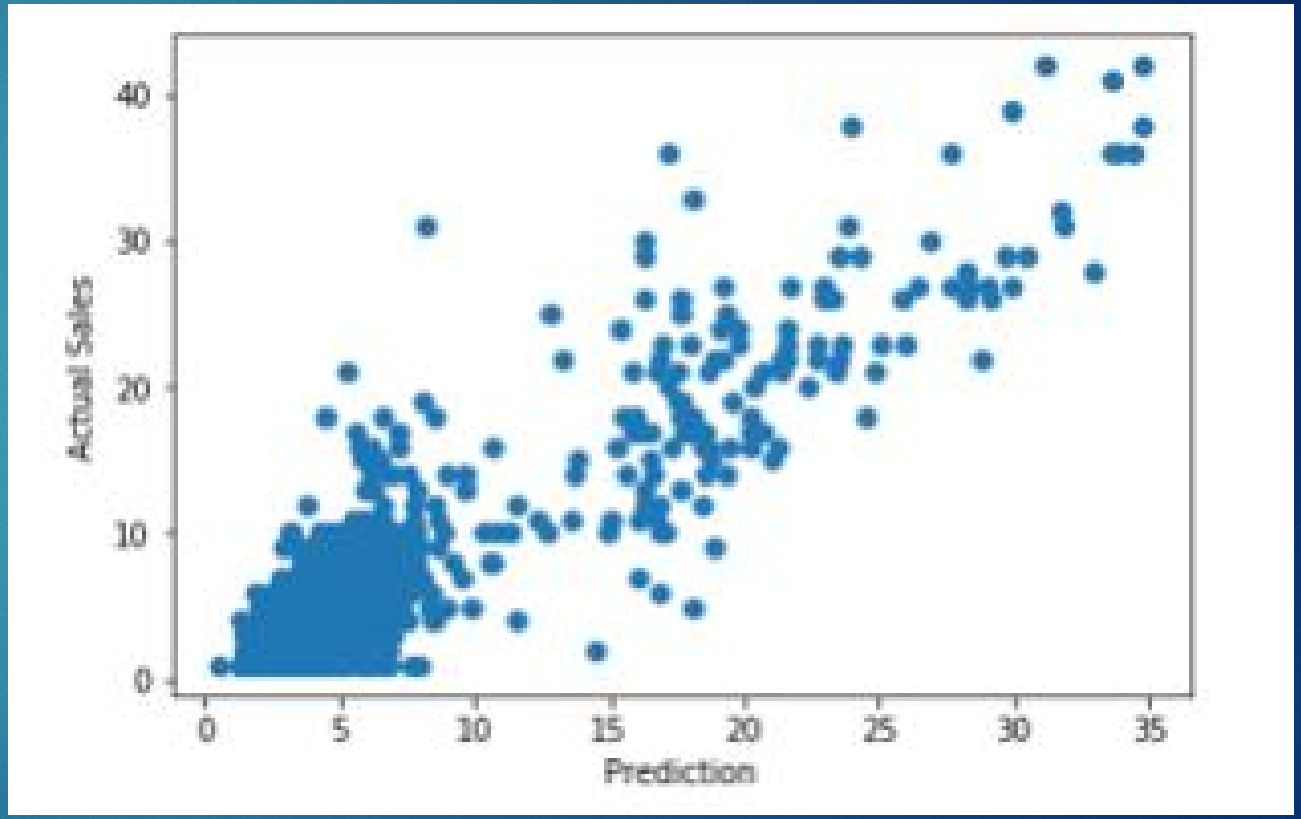
Random Forests Regression Model

- Score 0.90
- Mean Squared Error 3.965
- 5-fold CV – Score 0.676
- Best Params from GridSearch:
  - Max_depth 10
  - n_estimators 500
- Looks much more ‘linear’



Gradient Boosting Regressor Model

- Score 0.79
- Mean Squared Error 8.903
- 5-fold CV – Score 0.677
- Best Params from GridSearch:
  - Learing_rate 0.25
  - Max_depth 2
  - n_estimators 100
- These results are also linear but not as pronounced as the Random Forest Model



Test Data results – March 1-31

- Linear Regression model:
  - Score 0.65
  - Mean Squared Error 9.37
- Random Forest Regressor:
  - Score 0.74
  - Mean Squared Error 9.37
- Gradient Boosted Model:
  - Score 0.73
  - Mean Squared Error 9.98

Predicting how many baked items to make

- Ran prediction model with April 10, 2017 date (The next day)
- Used Random Forest Classifier
- Predication was rounded to the nearest whole number
- Limitations:
  - Predicting 'Seasonal' items
  - Is this right?

	Item_codes	Year	Month	Day	Weekday	Last_Week_Sold	2wks_Sold	pred	Item_name
0	1	2017	4	10	0	3	0	3.0	Alfajores
1	2	2017	4	10	0	0	0	12.0	Baguette
2	3	2017	4	10	0	15	7	17.0	Bread
3	4	2017	4	10	0	0	0	4.0	Bread Pudding
4	5	2017	4	10	0	0	0	4.0	Brownie
5	6	2017	4	10	0	5	0	5.0	Cake
6	7	2017	4	10	0	7	1	4.0	Cookies
7	8	2017	4	10	0	0	0	2.0	Empanadas
8	9	2017	4	10	0	0	0	1.0	Focaccia
9	10	2017	4	10	0	0	0	1.0	Frittata
10	11	2017	4	10	0	0	0	2.0	Fudge
11	12	2017	4	10	0	0	0	2.0	Kids biscuit
12	13	2017	4	10	0	0	0	3.0	Lemon and coconut
13	14	2017	4	10	0	3	0	3.0	Medialuna
14	15	2017	4	10	0	0	0	3.0	Muffin
15	16	2017	4	10	0	0	0	4.0	Panatone
16	17	2017	4	10	0	4	0	4.0	Pastry
17	18	2017	4	10	0	0	0	2.0	Pintxos
18	19	2017	4	10	0	1	0	3.0	Scandinavian
19	20	2017	4	10	0	5	3	3.0	Scone
20	21	2017	4	10	0	0	0	2.0	Tartine
21	22	2017	4	10	0	1	0	3.0	Tiffin
22	23	2017	4	10	0	4	3	3.0	Toast
23	24	2017	4	10	0	0	0	2.0	Vegan mincepie
24	25	2017	4	10	0	0	0	2.0	Victorian Sponge

Next Steps for this project

- More data!
  - One full year of sales data can account for all seasons/holidays
- Predict the rest of the items on the menu
- Start a database of ingredients and predict replenishment
- Add monetary data to predict P&L
- Look at more insights into what drives sales:
  - Bad weather?
  - Competition changes?
  - Experimentation with new products

Thank you!

- Chris Garland – Data Science Program at Thinkful
- Christopher.Garland@gmail.com

- Github - https://github.com/CascadiaRunner/Think_Capstone_2/blob/master/Supervised%20Learning%20-%20Capstone.ipynb