Forecasting Future Walmart Sales using KNN Regression

Muhammad Salman – 100995527 - Masters of Engineering in Electrical and Computer Engineering

Seyedeh Marziyeh Zamiri – 101214196 - Masters of Engineering in Electrical and Computer Engineering

Professor Mikhail Genkin, DATA 5000 Y – March 30th, 2021

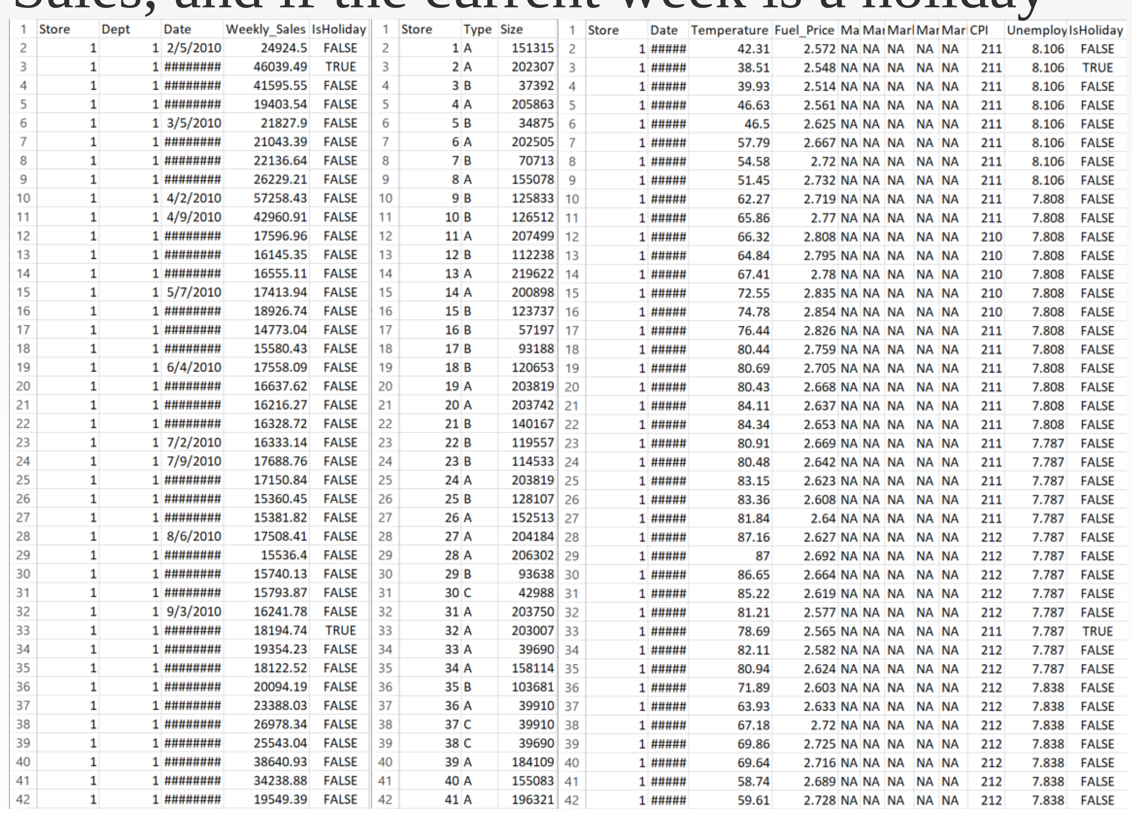
Introduction

- Walmart Data set imported from Kaggle
- Comprised of three main CSV files containing correlation of different features, based on data pulled from 45 different store locations, with a number of departments in each store
- Features vary from numerical values, such as amount of sales, to boolean True/False for if a week contains a holiday or not
- Goal is to build/train a model that can predict future Walmart sales based on a specific store/department and date

Figure 1: Sample Walmart Store

Dataset

- Relevant CSVs within the dataset are listed below, along with the features within each one
- *features.csv*: Store #, Date, Local Temperature, Local Fuel Price, Markdown/Discounts, CPI index, Unemployment Index, and if the current week is a holiday
- *stores.csv*: Store#, Type of Store, and size of store by number of products
- *train.csv*: Store#, Department#, Date, Weekly Sales, and if the current week is a holiday



1	Store	Dept	Date	Weekly_Sales	IsHoliday
2	1	1	2/5/2010	24924.5	FALSE
3	1	1	########	46039.49	TRUE
4	1	1	########	41595.55	FALSE
5	1	1	########	19403.54	FALSE
6	1	1	3/5/2010	21827.9	FALSE
7	1	1	########	21043.39	FALSE
8	1	1	########	22136.64	FALSE
9	1	1	########	26229.21	FALSE
10	1	1	4/2/2010	57258.43	FALSE
11	1	1	4/9/2010	42960.91	FALSE
12	1	1	########	17596.96	FALSE
13	1	1	########	16145.35	FALSE
14	1	1	########	16555.11	FALSE
15	1	1	5/7/2010	17413.94	FALSE
16	1	1	########	18926.74	FALSE
17	1	1	########	14773.04	FALSE
18	1	1	########	15580.43	FALSE
19	1	1	6/4/2010	17558.09	FALSE
20	1	1	########	16637.62	FALSE
21	1	1	########	16216.27	FALSE
22	1	1	########	16328.72	FALSE
23	1	1	7/2/2010	16333.14	FALSE
24	1	1	7/9/2010	17688.76	FALSE
25	1	1	########	17150.84	FALSE
26	1	1	########	15360.45	FALSE
27	1	1	########	15381.82	FALSE
28	1	1	8/6/2010	17508.41	FALSE
29	1	1	########	15536.4	FALSE
30	1	1	########	15740.13	FALSE
31	1	1	########	15793.87	FALSE
32	1	1	9/3/2010	16241.78	FALSE
33	1	1	########	18194.74	TRUE
34	1	1	########	19354.23	FALSE
35	1	1	########	18122.52	FALSE
36	1	1	########	20094.19	FALSE
37	1	1	########	23388.03	FALSE
38	1	1	########	26978.34	FALSE
39	1	1	########	25543.04	FALSE
40	1	1	########	38640.93	FALSE
41	1	1	########	34238.88	FALSE
42	1	1	########	19549.39	FALSE

1	Store	Type	Size
2	1	A	151315
3	2	A	202307
4	3	B	37392
5	4	A	205863
6	5	B	34875
7	6	A	202505
8	7	B	70713
9	8	A	155078
10	9	B	125833
11	10	B	126512
12	11	A	207499
13	12	B	112238
14	13	A	219622
15	14	A	200898
16	15	B	123737
17	16	B	57197
18	17	B	93188
19	18	B	120653
20	19	A	203819
21	20	A	203742
22	21	B	140167
23	22	B	119557
24	23	B	114533
25	24	A	203819
26	25	B	128107
27	26	A	152513
28	27	A	204184
29	28	A	206302
30	29	B	93638
31	30	C	42988
32	31	A	203750
33	32	A	203007
34	33	A	39690
35	34	A	158114
36	35	B	103681
37	36	A	39910
38	37	C	39910
39	38	C	39690
40	39	A	184109
41	40	A	155083
42	41	A	196321

1	Store	Date	Temperature	Fuel_Price	Ma	Mar	Marl	Mar	Mar	CPI	Unemploy	IsHoliday
2	1	#####	42.31	2.572	NA	NA	NA	NA	NA	211	8.106	FALSE
3	1	#####	38.51	2.548	NA	NA	NA	NA	NA	211	8.106	TRUE
4	1	#####	39.93	2.514	NA	NA	NA	NA	NA	211	8.106	FALSE
5	1	#####	46.63	2.561	NA	NA	NA	NA	NA	211	8.106	FALSE
6	1	#####	46.5	2.625	NA	NA	NA	NA	NA	211	8.106	FALSE
7	1	#####	57.79	2.667	NA	NA	NA	NA	NA	211	8.106	FALSE
8	1	#####	54.58	2.72	NA	NA	NA	NA	NA	211	8.106	FALSE
9	1	#####	51.45	2.732	NA	NA	NA	NA	NA	211	8.106	FALSE
10	1	#####	62.27	2.719	NA	NA	NA	NA	NA	211	7.808	FALSE
11	1	#####	65.86	2.77	NA	NA	NA	NA	NA	211	7.808	FALSE
12	1	#####	66.32	2.808	NA	NA	NA	NA	NA	210	7.808	FALSE
13	1	#####	64.84	2.795	NA	NA	NA	NA	NA	210	7.808	FALSE
14	1	#####	67.41	2.78	NA	NA	NA	NA	NA	210	7.808	FALSE
15	1	#####	72.55	2.835	NA	NA	NA	NA	NA	210	7.808	FALSE
16	1	#####	74.78	2.854	NA	NA	NA	NA	NA	210	7.808	FALSE
17	1	#####	76.44	2.826	NA	NA	NA	NA	NA	211	7.808	FALSE
18	1	#####	80.44	2.759	NA	NA	NA	NA	NA	211	7.808	FALSE
19	1	#####	80.69	2.705	NA	NA	NA	NA	NA	211	7.808	FALSE
20	1	#####	80.43	2.668	NA	NA	NA	NA	NA	211	7.808	FALSE
21	1	#####	84.11	2.637	NA	NA	NA	NA	NA	211	7.808	FALSE
22	1	#####	84.34	2.653	NA	NA	NA	NA	NA	211	7.808	FALSE
23	1	#####	80.91	2.669	NA	NA	NA	NA	NA	211	7.787	FALSE
24	1	#####	80.48	2.642	NA	NA	NA	NA	NA	211	7.787	FALSE
25	1	#####	83.15	2.623	NA	NA	NA	NA	NA	211	7.787	FALSE
26	1	#####	83.36	2.608	NA	NA	NA	NA	NA	211	7.787	FALSE
27	1	#####	81.84	2.64	NA	NA	NA	NA	NA	211	7.787	FALSE
28	1	#####	87.16	2.627	NA	NA	NA	NA	NA	212	7.787	FALSE
29	1	#####	87	2.692	NA	NA	NA	NA	NA	212	7.787	FALSE
30	1	#####	86.65	2.664	NA	NA	NA	NA	NA	212	7.787	FALSE
31	1	#####	85.22	2.619	NA	NA	NA	NA	NA	212	7.787	FALSE
32	1	#####	81.21	2.577	NA	NA	NA	NA	NA	212	7.787	FALSE
33	1	#####	78.69	2.565	NA	NA	NA	NA	NA	211	7.787	TRUE
34	1	#####	82.11	2.582	NA	NA	NA	NA	NA	212	7.787	FALSE
35	1	#####	80.94	2.624	NA	NA	NA	NA	NA	212	7.787	FALSE
36	1	#####	71.89	2.603	NA	NA	NA	NA	NA	212	7.838	FALSE
37	1	#####	63.93	2.633	NA	NA	NA	NA	NA	212	7.838	FALSE
38	1	#####	67.18	2.72	NA	NA	NA	NA	NA	212	7.838	FALSE
39	1	#####	69.86	2.725	NA	NA	NA	NA	NA	212	7.838	FALSE
40	1	#####	69.64	2.716	NA	NA	NA	NA	NA	212	7.838	FALSE
41	1	#####	58.74	2.689	NA	NA	NA	NA	NA	212	7.838	FALSE
42	1	#####	59.61	2.728	NA	NA	NA	NA	NA	212	7.838	FALSE

Figure 2: Snapshot of Three CSV files, Train, Stores, Features (Left to Right)

Methodology

Data Cleaning

- Once data was imported using Excel, all following data analysis was done using Python
- The first step was to load the data using the Pandas package, and import the CSV table as a data frame
- We proceeded to remove N/A data, as can be seen in the CSV screenshots in the Dataset section
- Once completed, all three data frames were merged into one, named main
- The main data frame was created by merging on common features between the three data frames, such as Store# and Date

```
#Load all excel files in through pandas
train  = pandas.read_csv("train.csv")
features = pandas.read_csv("features.csv")
stores  = pandas.read_csv("stores.csv")

#Some values maybe NA depending on stores
#Clean up data

#For unavaialble markdowns, fill with 0
features['Markdown1'] = features["MarkDown1"].fillna(0)
features['Markdown2'] = features['MarkDown2'].fillna(0)
features['Markdown3'] = features['MarkDown3'].fillna(0)
features['Markdown4'] = features['MarkDown4'].fillna(0)
features['Markdown5'] = features['MarkDown5'].fillna(0)
```

Figure 3: Code to Load CSV and Clean Data

Data Exploration

- The next step was to explore the data and observe strong correlation between features
- The three sets that were observed were: Type of Store vs. Weekly Sales, Type of Store vs. Size of Store, Is Holiday vs. Weekly Sales

Observations

- Type of Store vs. Weekly Sales: Weak correlation, Type C seems to have least amount of sales, no strong difference between A and B
- Type of Store vs. Size of Store: Strong Correlation, A is largest, B is second largest, C is smallest
- Is Holiday vs. Weekly Sales: Some correlation, with slightly more sales when there was a holiday, however not as much as expected

Figure 4: Type of Store vs. Size of Store

Data Modelling

- In order to model the data, a machine learning model is normally selected
- We will select the K Nearest Neighbors Regression model
- Regression models allow a solid value to be predicted based on a series of dependent variables, such as we have in our case, since we are predicting future sales per store and department at certain dates
- The KNN approach will allow features to be classified based on their similarities as described with input data. Hence, predicted values will be classified based on how much their data resembles points in the dataset
- This modelling can be done in python using scikit-learn library in Python
- *Yet to be finalized*

Results

- *No final results currently yielded*
- *Mention if trained model is successful*

Conclusion

- *Concluding remarks*
- *Explain how this model could help predict and forecast future sales for department stores*

References

[1] Kaggle.com. 2014. Walmart Recruiting - Store Sales Forecasting | Kaggle. [online] Available at: https://www.kaggle.com/c/walmart-recruiting-store-sales-forecasting/data [Accessed 31 January 2021].

Acknowledgments

We would like to thank the organizers of Data Day for allowing us to present our Walmart Forecaster Project, as a part of the DATA 5000 course.

We would also like to thank Professor Michael Genkin for his dedication and support during the semester, as well as Professor Komeili and Professor Velazquez for teaching the course.