

Sales Forecasting of Woody's Cafe

Group members: Faheem Kamaludeen, Nishitha Chidipothu , Dillirajan Sankar



Introduction

“Sales Forecasting is the process of using a company’s sales records over the past years to predict the short-term or long-term sales performance of that company in the future. This is one of the pillars of proper financial planning.” But on a general note risk and uncertainty is unavoidable everywhere.

Sales forecast is considered to be the main idea for creating your future plans for the business. And based on sales forecast output the profit and loss statement , cash flow statement and balance sheet can be prepared. The steps to create a sales forecast will be to list the items that we want to forecast and are being sold. Either estimate the sales that might happen or refer to the previous data that's available. And implement the models on the cleaned data and compare the available and forecasted one to plan out the business model.

Background/Motivation

Despite the many advantages of using a sales forecasting model, companies are hesitant to try machine learning models because of the risk of using an inaccurate model. Research shows that only 55% of the sales leaders have high confidence in their models. Several factors can affect the sales of business and several which can give a perception of affecting the sales. It is important to find both the factors affecting sales and also a model with good accuracy to be confident in sales forecasting. Here, we are performing sales forecasting for a small café, Woody's cafe in Rutgers University using various machine-learning models and comparing the results by using different evaluation metrics.

Concept to be Learned

The concept to be learned is the forecasting of sales for Woody's cafe. Sales forecasting would be performed in terms of both volume of sales and the total sales. To predict sales forecasting for woody's café, supervised learning models like regression and time-series forecasting models will be used.

Obtaining and Cleaning Data

- Data obtained from Woody's Cafe, provided by the manager, Ms. Tiffany Moon.
- Split of date and time into separate columns and fill the empty spaces with date and time respectively.
- Conversion of irrelevant empty columns and rows into null, followed by removal of such null instances.
- Aggregation of data based on "Date" to get one unique date in each instance.
- Addition of external factors namely weather, season, Food Consumer Price Index (CPI), Holiday week, and Break week
- Implemented one-hot encoding on weather attribute and ordinal encoding on season attribute.

Sales, weather, and CPI Data (To be Merged)

Date/Time Period	Item Description	Item Number	Qty	Avg Item Price	Item Sales Total
01/04/2022					
07:00 AM - 07:59 AM	Breakfast Potatoes	457	1	2.5	2.5
07:00 AM - 07:59 AM	Hourly Item Sales Totals:		1		2.5
08:00 AM - 08:59 AM	12oz Hot Mug Refill	301	1	2	2
	French Toast Platter	310	1	1.25	1.25
	Breakfast Potatoes	365	1	5.5	5.5
	Pineapple Cups	457	1	2.5	2.5
		6012	1	3.25	3.25
08:00 AM - 08:59 AM	Hourly Item Sales Totals:		5		14.5
09:00 AM - 09:59 AM	Egg Meat and Cheese	406	5	5	25
	12oz Hot	301	2	2	4
	Small Hot Tea	303	1	1.75	1.75
	Fruit	852	1	1	1
09:00 AM - 09:59 AM	Hourly Item Sales Totals:		9		31.75

18	2020	M06	2020 Jun	284.462
19	2020	M07	2020 Jul	282.301
20	2020	M08	2020 Aug	282.440
21	2020	M09	2020 Sep	282.165
22	2020	M10	2020 Oct	283.504
23	2020	M11	2020 Nov	283.049
24	2020	M12	2020 Dec	283.892
25	2021	M01	2021 Jan	283.787
26	2021	M02	2021 Feb	284.173
27	2021	M03	2021 Mar	285.064
28	2021	M04	2021 Apr	287.250
29	2021	M05	2021 May	287.277
30	2021	M06	2021 Jun	289.528
31	2021	M07	2021 Jul	291.792
32	2021	M08	2021 Aug	293.132
33	2021	M09	2021 Sep	297.658
34	2021	M10	2021 Oct	299.049
35	2021	M11	2021 Nov	301.334
36	2021	M12	2021 Dec	302.121
37	2022	M01	2022 Jan	305.120
38	2022	M02	2022 Feb	307.079

name	datetime	tempmax	tempmin	temp	feelslikem	feelslikei	feelslike	dew	humidity	precip	precipprot	precipcovr	preciptype	snow	snowdepth	windgust	windspeed	windirr	sealevelpr	cloudcovr	visibility	solarradiat	so
New Bruns	7/1/2021	84.1	72.1	76.1	87.7	72.1	76.9	70.3	83.32	0.64		45.83		0	0	25.3	9	243.2	1010.6	68	7.1	153.9	
New Bruns	7/2/2021	81	64.7	73.4	83.2	64.7	73.7	66.7	80.42	0.47		45.83		0	0	21.9	9	210.9	1006.5	72.8	8.5	249.3	
New Bruns	7/3/2021	68.6	61.5	64.9	68.6	61.5	64.9	58.7	80.52	0.06		29.17		0	0	24.2	11.7	42.4	1008.4	99.4	9.2	79.9	
New Bruns	7/4/2021	81.5	62.1	71.1	81.4	62.1	71.2	59.6	69.77	0		4.17		0	0	9.2	168.4	1011.6	67.7	8.8	184		
New Bruns	7/5/2021	86.6	63.3	75.8	87.4	63.3	76.3	63.5	68.26	0		0		0	0	21.9	7.9	155.6	1016.1	21.5	7	414.4	
New Bruns	7/6/2021	94.2	68.3	79.3	99.8	68.3	81.5	67.7	70.04	0.86		20.83		0	0	37.8	17	244.2	1014.5	33.3	7.8	221.1	
New Bruns	7/7/2021	93.4	69.7	81.8	99.2	69.7	85.4	70.4	71.43	0.02		8.33		0	0	21.9	7.9	259.3	1013.5	27.6	7.5	257.3	
New Bruns	7/8/2021	87.4	72.8	79.2	94.7	72.8	81.4	70.8	75.85	0.2		25		0	0	20.8	9.5	131.8	1013.8	58	7.8	188.9	
New Bruns	7/9/2021	87.3	72	77.9	91.2	72	79.3	69.5	77.19	1.42		54.17		0	0	31.1	14.3	236.8	1009.4	77.1	7.8	229.1	
New Bruns#####	83.6	68	75.4	85.3	68	75.8	66.8	76.12	0		0	0		0	0	11	182.5	1014.1	57.6	8.5	205.1		
New Bruns#####	82	71.5	75.5	85.2	71.5	75.9	69.1	81.15	0.03		8.33		0	0	0	20.8	10.5	121.7	1017.9	75.6	8.8	169.8	
New Bruns#####	89.8	72.1	78.7	99.4	72.1	81.3	72.6	82.61	0.58		29.17		0	0	0	8.5	178.9	1018.4	79.5	7.8	133.2		
New Bruns#####	76	70.4	72.9	76	70.4	72.9	69.4	88.75	0.03		16.67		0	0	0	7.5	98.7	1022.4	100	6.5	98.1		
New Bruns#####	88.9	71.4	77.9	94.3	71.4	79.6	70.5	79.6	0.09		8.33		0	0	0	32.2	10.5	149.5	1020.3	66	6.4	240	
New Bruns#####	89	70.6	80.5	94.6	70.6	83	69.5	70.88	0		0	0		0	0	9.2	218	1017.8	33.8	7.8	204.1		
New Bruns#####	92.7	74	83.4	97.8	74	86	70	66.56	0		0	0		0	0	20.8	8.3	239.1	1015.4	23.5	8.1	227.4	
New Bruns#####	92.3	72.7	79.8	104.7	72.7	82.6	71.3	76.02	1.36		37.5		0	0	0	57.5	14.8	172.4	1015.8	45.2	7.2	144.2	

Attributes

Before Cleaning

Date/Time Period
Item Description
Item Number
Quantity
Average item price
Item sales total

After Cleaning

Date
Holiday Week (0, 1)
Break Week (0,1)
Consumer Price Index (float)
Clear-day
Cloudy
Partly-Cloudy-Day
Rain
Snow
Season (0,1, 2)
Quantity
Item Sales Total

Ideal Dataset


	Date	Quantity	CPI	Holiday week	Break week	clear-day	cloudy	partly-cloudy-day	rain	snow	Season	Item Sales Total
0	2021-07-19	68.0	291.792	0	0	0.0	0.0	1.0	0.0	0.0	1.0	163.16
1	2021-07-20	85.0	291.792	0	0	0.0	0.0	1.0	0.0	0.0	1.0	263.04
2	2021-07-21	94.0	291.792	0	0	0.0	0.0	0.0	1.0	0.0	1.0	237.92
3	2021-07-22	123.0	291.792	0	0	0.0	0.0	1.0	0.0	0.0	1.0	376.07
4	2021-07-23	152.0	291.792	0	0	0.0	0.0	1.0	0.0	0.0	1.0	425.73
...
149	2022-03-02	2574.0	305.024	0	0	0.0	0.0	1.0	0.0	0.0	2.0	10450.58
150	2022-03-03	2503.0	305.024	0	0	0.0	0.0	0.0	1.0	0.0	2.0	10653.23
151	2022-03-04	1746.0	305.024	0	0	1.0	0.0	0.0	0.0	0.0	2.0	6715.19
152	2022-03-07	2655.0	305.024	0	0	0.0	0.0	0.0	1.0	0.0	2.0	10739.45
153	2022-03-08	2720.0	305.024	0	0	0.0	0.0	1.0	0.0	0.0	2.0	10944.55


154 rows × 12 columns


Sliding Window Technique- Cross Validation

{1, 2, 3, 4, 5, 6, 7, 8, 9, 10}

{1, 2, 3, 4, 5, 6, 7 | 8, 9, 10}


{1, 2, 3, 4, 5, 6, 7 | 8, 9, 10}


{1, 2, 3, 4, 5, 6, 7 | 8, 9, 10}


{1, 2, 3, 4, 5, 6, 7 | 8, 9, 10}

Sliding Window Technique

- For our project, we have 154 instances of data (154 days).
- We have split the data into 107 instances for training and 47 instances for testing.
- The first iteration of sliding window trains the first 107 instances (Instances 1 - 107) and performs prediction on the first 5 test instances (Instances 108 - 112).
- The values of evaluation metrics are determined for every iteration.
- The second iteration of sliding window moves one instance forward and trains the next 107 instances (Instances 2 - 108) and performs prediction on the next 5 test instances (Instances 109 - 113).
- Likewise, prediction is done till the entire dataset is over. In our case, there were 42 iterations of sliding window.
- With 42 sets of evaluation metric values from 42 iterations, we take the average of these evaluation metric.

Approach and Evaluation Metrics

Algorithms used (Supervised)-

- Random Forest Regression
- Extra Trees Regressor
- Support Vector Regression
- LG Boost
- Multiple Linear Regression

Gaussian Naive Bayes as
Baseline

Evaluation metrics used-

- Mean Absolute Error
- Mean Absolute Percentage Error
- Root Mean Squared Error



Evaluation Metric Comparison

metrics_q

	Models	MAE	MAPE	RMSE	P-Value
0	Baseline Gaussian Naive Bayes	1108.957124	0.611715	1220.207661	0.046512
1	Multiple Linear Regression	93.655666	0.142383	110.733419	0.021021
2	Random Forest Regression	152.054153	0.107496	189.682569	0.017931
3	LGBost	290.752783	0.219329	346.503927	0.027834
4	Extra Trees Regressor	258.987368	0.224827	312.742493	0.018751
5	Support Vector Regression	93.293529	0.113159	110.467236	0.020534

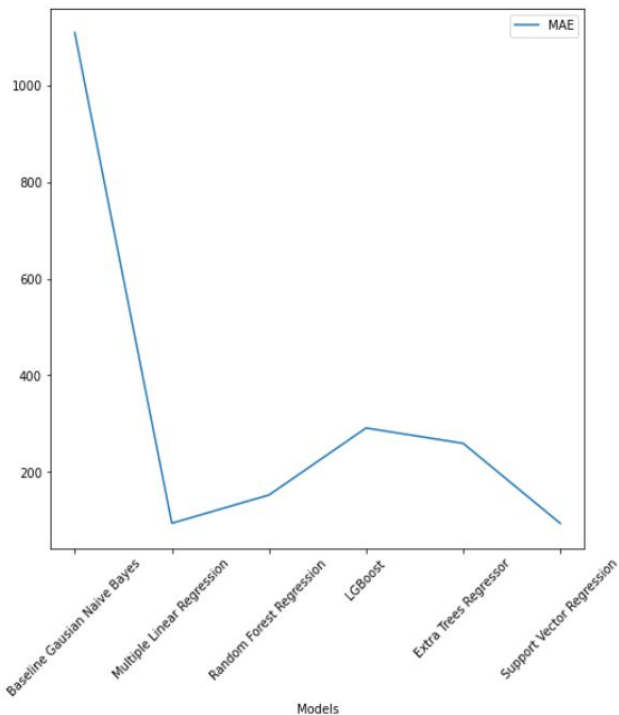
metrics_s

	Models	MAE	MAPE	RMSE	P-Value
0	Baseline Gaussian Naive Bayes	3476.805333	0.537930	3838.501895	0.046512
1	Multiple Linear Regression	356.823149	0.096490	443.499731	0.026094
2	Random Forest Regression	573.897727	0.099361	707.454895	0.032465
3	LGBost	1252.759527	0.248177	1451.268104	0.038907
4	Extra Trees Regressor	1246.666012	0.263588	1438.590652	0.030801
5	Support Vector Regression	339.181212	0.147509	407.968249	0.026458

Error Plots for each Model (Prediction of Quantity)

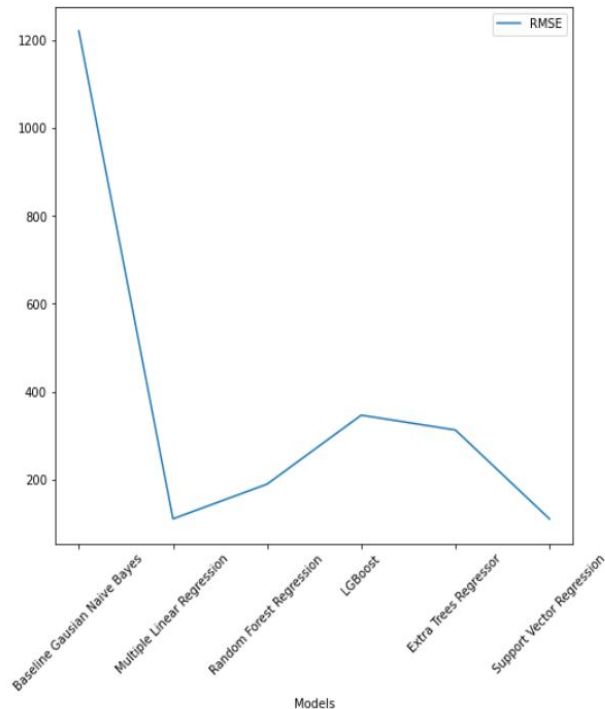
```
metrics_q.plot(x='Models', y='MAE', kind='line', figsize=(8,8), rot=45)
```

<AxesSubplot:xlabel='Models'>

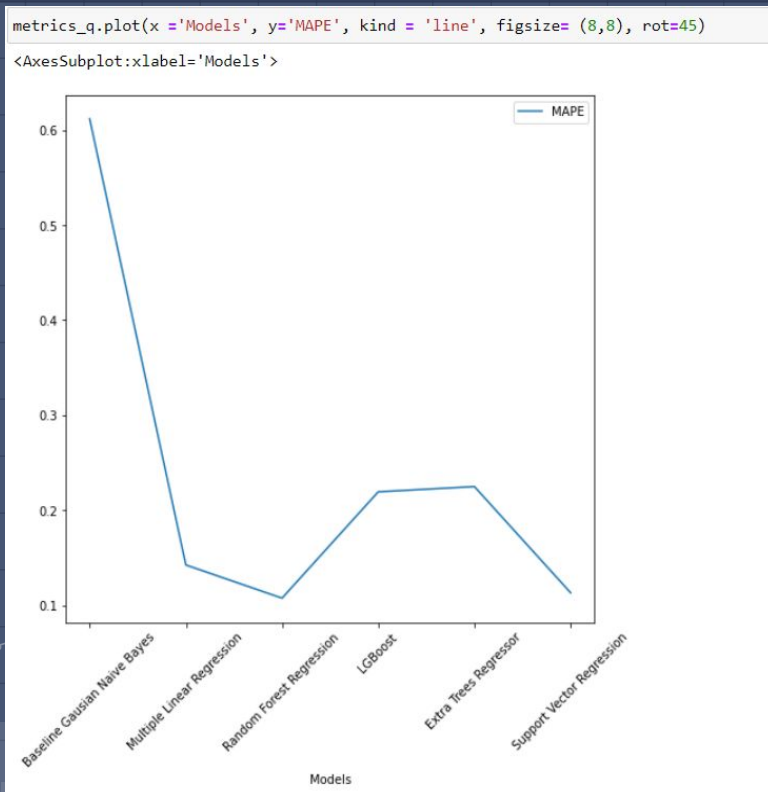


```
metrics_q.plot(x='Models', y='RMSE', kind='line', figsize=(8,8), rot=45)
```

<AxesSubplot:xlabel='Models'>



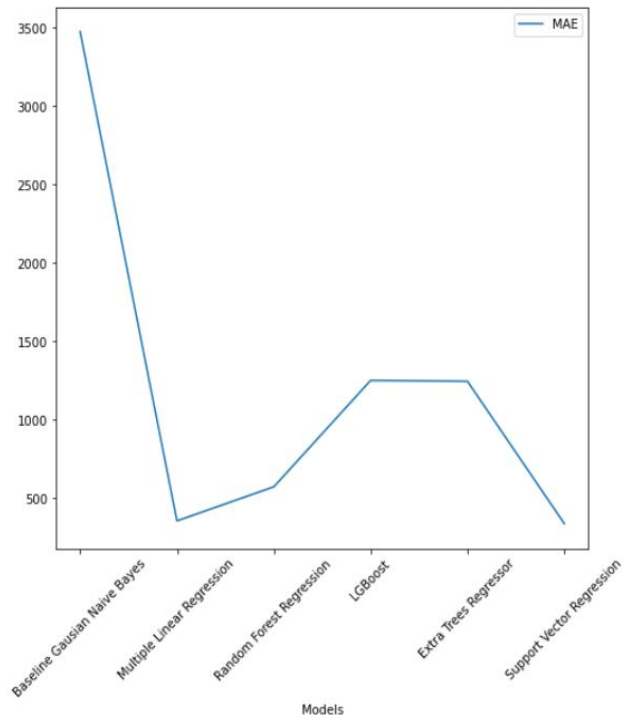
Error Plots for each Model (Prediction of Quantity)



Error Plots for each Model (Prediction of Item Sales Total)

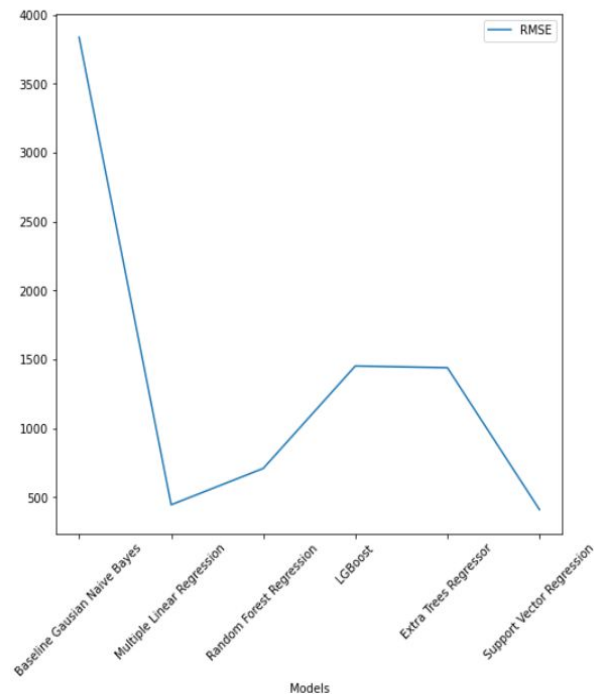
```
metrics_s.plot(x='Models', y='MAE', kind='line', figsize=(8,8), rot=45)
```

<AxesSubplot:xlabel='Models'>



```
metrics_s.plot(x='Models', y='RMSE', kind='line', figsize=(8,8), rot=45)
```

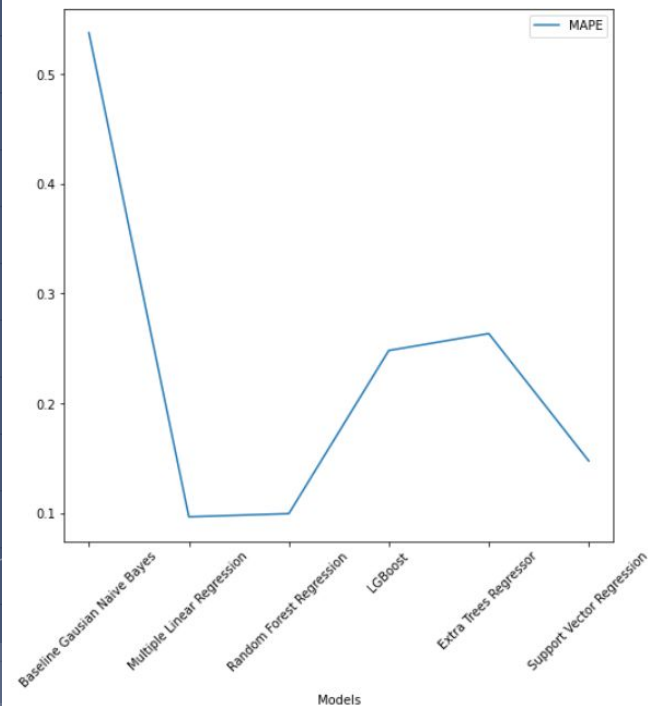
<AxesSubplot:xlabel='Models'>



Error Plots for each Model (Prediction of Item Sales Total)

```
metrics_s.plot(x='Models', y='MAPE', kind='line', figsize=(8,8), rot=45)
```

<AxesSubplot:xlabel='Models'>



Discussion and Conclusions

- The baseline used is Gaussian Naives Bayes.
- We have trained and tested our dataset on 6 models apart from the baseline and evaluated using 3 evaluation metrics namely MSE, MAPE, and RMSE.
- The results showed that the best model under all these conditions is Random Forest Regression.
- Support Vector Regression is working best for predicting both the sales volume and Item sales total.
- The second best model is Random Forest Regressor.
- LGBost and Extra Tree Regressor perform worse on predicting both the sales volume and the item sales total.

Future Work/Improvements

- Hyper-parameter tuning for better performance of the models.
- More Data Instances for better training and performance of the model.
- Sales forecasting for individual items could be implemented.

Literature Review

1. K. Saraswathi, N. T. Renukadevi, S. Nandhinidevi, S. Gayathridevi, and P. Naveen, "Sales prediction using machine learning approaches", AIP Conference Proceedings 2387, 140038 (2021), DOI: 10.1063/5.0068655
2. Robert Fildes, Stuart Bretschneider, Fred Collopy, Michael Lawrence, Doug Stewart, Heidi Winklhofer, John T. Mentzer, Mark A. Moon, "Researching Sales Forecasting Practice: Commentaries and authors' response on "Conducting a Sales Forecasting Audit" by M.A. Moon, J.T. Mentzer & C.D. Smith", International Journal of Forecasting, 19:1 (2003), pp. 27-42, DOI: 10.1016/S0169-2070(02)00033-X.
3. Y. Niu, "Walmart Sales Forecasting using XGBoost algorithm and Feature engineering," 2020 *International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE)*, 2020, pp. 458-461, DOI: 10.1109/ICBASE51474.2020.00103.
4. Samaneh Beheshti-Kashi, Hamid Reza Karimi, Klaus-Dieter Thoben, Michael Lütjen & Michael Teucke (2015) "A survey on retail sales forecasting and prediction in fashion markets", Systems Science & Control Engineering, 3:1, 154-161, DOI: 10.1080/21642583.2014.999389.
5. Carina Intan Permatasari, Wahyudi Sutopo, and Muh. Hisjam, "Sales forecasting newspaper with ARIMA: A case study", AIP Conference Proceedings 1931, 030017 (2018), DOI: 10.1063/1.5024076
6. Pavlyshenko, B.M., "Machine-Learning Models for Sales Time Series Forecasting.", Data (2019) 4:1, 15, DOI: 10.3390/data4010015

Project Time Log

Dillirajan Sankar

2/28:3-7:4: Project Ideas-Brainstorming as group
 2/29:11.00-1.00:2: Dataset Search
 3/03:9.00-9.30:0.5: Discussion of project ideas with Dr.Bryan
 3/05:10.00-1.00:3: Research on Sales forecasting and Stock Fund Price Forecasting
 3/07:9.30-10.30:1: Brainstorming project ideas and feasibility with Dr. Tingting
 3/08:4-5:1: Meet with Cafe's manager, Ms Tiffany Moon
 3/08:5-6:1: Project discussion as group
 3/09:6-6.30:0.5: Ideas feasibility meeting with Dr. William Pottenger
 3/09:9.30-10.30:1: Project discussion as group
 3/10:10-12:2: Research on regression and time-series forecasting models
 3/11:10-11:2: Exploration of woody's cafe sales data
 3/14:11-1:2: Research on cleaning to be done on sales data and writing
 3/15:11-12:1: Study of evaluation metrics for sales forecasting
 3/16:11-1:2: Concept description and approach research and writing
 3/19:1-4:3: Group meeting to discuss and explain individual parts to each other
 3/19:5-6:1: Discussion with Mrs. Vanesa (Sales Forecasting Expert) for some insights
 3/20:6-7:1: Proof reading of report and presentation slides (28 hours till now)
 4/01:4-6:2: Discussion on all steps required for Data Cleaning
 4/02:12-2:2: Data Cleaning- Strategy for separating date and time into separate columns
 4/04:2-5:3: Data Cleaning- Separating Time from Date and filling in empty space with respective time periods
 4/15:5.30-6.15:0.75: Project Queries- Sliding Window
 4/22:12-2:2: Project discussion with group and worked on different attributes and the data that should be added
 4/22:3-7:4: Implemented addition of Holiday week , break week and weather attributes along with one-hot encoding implementation on weather data
 4/23:12-2:2: Implementation code for the various ML and Deep Learning models
 4/23:3-7:4: Learning of various concepts of cross-validation for Time-series forecasting
 4/25:9-9.30:0.5: Project Queries- Sliding Window
 4/25:8-10:2: Discussion with the team on aggregation of data and debugging
 4/26:4-7.30:3.5: Implementation of Sliding Window Cross-Validation
 4/26:8.45-11.45:3: Implementation of ETR, SVR, XGBoost, and LGBBoost models and Debugging
 4/26:12-12.30:0.5: Learning of plotting Time-Series Graph
 4/27:11-3:4: Project Presentation and compilation
 Total Hours- 60.25 hours

Faheem Kamaludeen Mohideen

2/26:4-6:2: Datasets search
 2/28:3-7:4: Ideas Brainstorming as group
 3/01:4-8:4: Sentiment analysis research for data and models
 3/03:4-6:2: Trying feasibility of different ideas
 3/03:9-9.30:0.5: Discussion about the ideas with Dr.Bryan
 3/07:9.30-10.30:1: Ideas feasibility meeting with Dr. Tingting
 3/08:4-5:1: Meet with Cafe's manager
 3/08:5-6:1: Project discussion as group
 3/09:6-6.30:0.5: Ideas feasibility meeting with Dr. Bill
 3/09:9.30-10.30:1: Project discussion as group
 3/12:3.30-5.30:2: Exploration of data
 3/13:6-8:2: Project proposal - Introduction writing
 3/14:3-7.30:4.5: Literature review exploration and writing
 3/19:1-4:3: Group meeting to discuss and explain individual parts to each other
 3/20:6-7:1: Proof reading of report and presentation slides
 3/28:6-9:3: Initial Data Cleaning
 4/01:4-6:2: Discussion on all steps required for Data Cleaning
 4/02:12-2:2: Data Cleaning- Strategy for separating date and time into separate columns
 4/15:5.30-6.15:0.75: Project Queries- Sliding Window
 4/22:12-2:2: Project discussion with group and worked on different attributes and the data that should be added
 4/22:3-7:4: Implemented addition of Holiday week , break week and weather attributes along with one-hot encoding implementation on weather data
 4/23:5-9:4: Data Cleaning Debugging
 4/24:6-8:2: Seasons implementation
 4/25:9-9.30:0.5: Project Queries- Sliding Window
 4/25:8-10:2: Discussion with the team on aggregation of data and debugging
 4/26:4-7.30:3.5: Implementation of Sliding Window Cross-Validation
 4/26:8.45-11.45:3: Debugging
 4/27:11-3:4: Implementing Plots and Project Presentation discussion

Total: 62.25 Hours

Nishitha Chidipothu

2/28:3-7:4: Ideas Brainstorming as group
3/03:9.00-9.30:0.5: Discussion about the ideas with Dr. Bryan
3/07:9.30-10.30:1: Ideas feasibility meeting with Dr. Tingting
3/08:4-5:1: Meet with Cafe's manager
3/08:5-6:1: Project discussion as group
3/09:11-2:3: Referring to ideas given by Dr. Tingting and exploring new ideas
3/09:6-6.30:0.5: Ideas feasibility meeting with Dr. Bill
3/09:9.30-10.30:1: Project discussion as group
3/11:10-12:2: Exploration of the data given by the cafe manager
3/13:4-6:2: Usage of different attributes understanding and writing for project proposal
3/16:1-4:3: Step by step process for evaluation and presentation slides.
3/19:10.30-11/30:1: Understanding different evaluation metrics
3/19:5-6:1: Discussion with Vanesa (Sales forecasting expert) for some insights
3/19:1-4:3: Group meeting to discuss and explain individual parts to each other
3/20:6-7:1: Proof reading of report and presentation slides
4/1:4-6:2: Discussion on steps with respect to data cleaning
4/6:7-10:3: Split date into separate columns and fill the empty spaces with date respectively.
4/8:7-10:3: Split time-range into separate columns and fill the empty spaces with time-range respectively.
4/15:5.30-6.15:0.75: Discussion on sliding window with Dr. Tingting
4/22-11-2:3: Project discussion with group and worked on different attributes and the data that should be added
4/22-3-7:4: Implemented addition of Holiday week, break week and weather attributes.
4/23-5-8:3: Implemented one-hot encoding on Weather attribute and checked on debugging
4/25-9-9.30:0.5: Doubts clearing session with Dr. Tingting
4/25-8-10 pm:2: Discussion with the team on aggregation of data and debugging
4/26-5.30-6.30pm:1: Working on encoding of season attribute and implementation of sliding window
4/27:11-3:4: Implementing Plots and Project Presentation discussion
4/27:4-6:2: Preparing for the presentation

Total hours:58.5 hours



THANK YOU!