Data Science Finhack2

Anushree Tomar

Problem statement

LTFS receives a lot of requests for its various finance offerings that include housing loan, two-wheeler loan, real estate financing, and microloans. The number of applications received is something that varies a lot with the season. Going through these applications is a manual process and is tedious. Accurately forecasting the number of cases received can help with resource and manpower management resulting into quick response on applications and more efficient processing. You have been appointed with the task of forecasting daily cases for the next 3 months for 2 different business segments at the country level keeping in consideration the following major Indian festivals (inclusive but not exhaustive list): Diwali, Dussehra, Ganesh Chaturthi, Navratri, Holi, etc

Data Dictionary

The train data has been provided in the following way:

- For business segment 1, historical data has been made available at branch ID level
- For business segment 2, historical data has been made available at the State level.

Exploratory Data Analysis

Train data

```
##    application_date segment branch_id       state zone case_count
## 1:       2017-04-01       1         1 WEST BENGAL EAST         40
## 2:       2017-04-03       1         1 WEST BENGAL EAST          5
## 3:       2017-04-04       1         1 WEST BENGAL EAST          4
## 4:       2017-04-05       1         1 WEST BENGAL EAST        113
## 5:       2017-04-07       1         1 WEST BENGAL EAST         76
## 6:       2017-04-12       1         1 WEST BENGAL EAST        123
```

```
##    application_date segment branch_id       state zone case_count
## 1:       2019-07-18       2        NA WEST BENGAL <NA>       2408
## 2:       2019-07-19       2        NA WEST BENGAL <NA>       1886
## 3:       2019-07-20       2        NA WEST BENGAL <NA>       1480
## 4:       2019-07-21       2        NA WEST BENGAL <NA>       1028
## 5:       2019-07-22       2        NA WEST BENGAL <NA>       1946
## 6:       2019-07-23       2        NA WEST BENGAL <NA>       1984
```

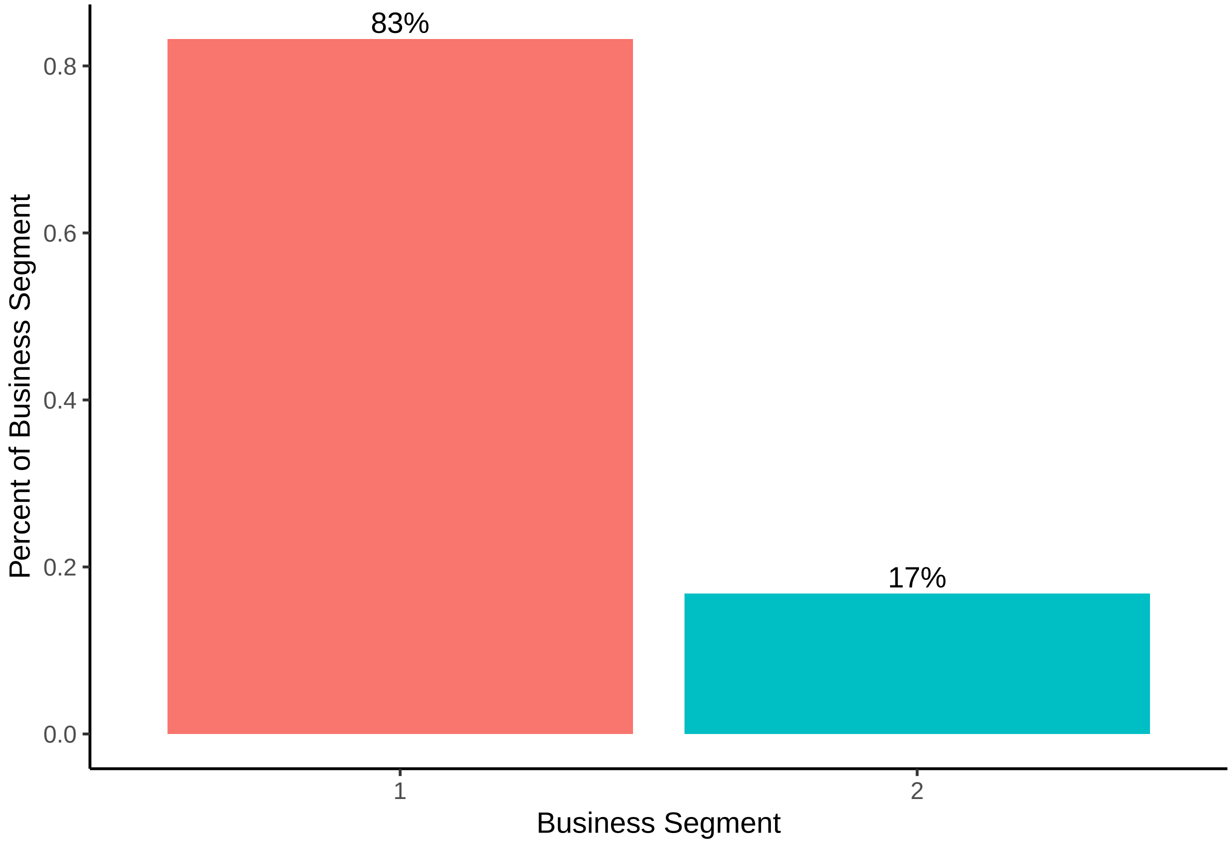
There is an “NA” in zone column. We can identify zone by analyzing combination of state and zone column.

```
##              state    zone
##  1:          ASSAM    EAST
##  2:          BIHAR    EAST
##  3:   CHHATTISGARH CENTRAL
##  4:        GUJARAT    WEST
##  5:        HARYANA   NORTH
##  6:      JHARKHAND   SOUTH
##  7:         KERALA   SOUTH
##  8:      KARNATAKA   SOUTH
##  9:    MAHARASHTRA    WEST
## 10: MADHYA PRADESH CENTRAL
## 11:         ORISSA   SOUTH
## 12:         PUNJAB   NORTH
## 13:     TAMIL NADU   SOUTH
## 14:        TRIPURA    EAST
## 15:  UTTAR PRADESH    EAST
## 16:    WEST BENGAL    EAST
```

Above list is the unique list of state and missing zone.

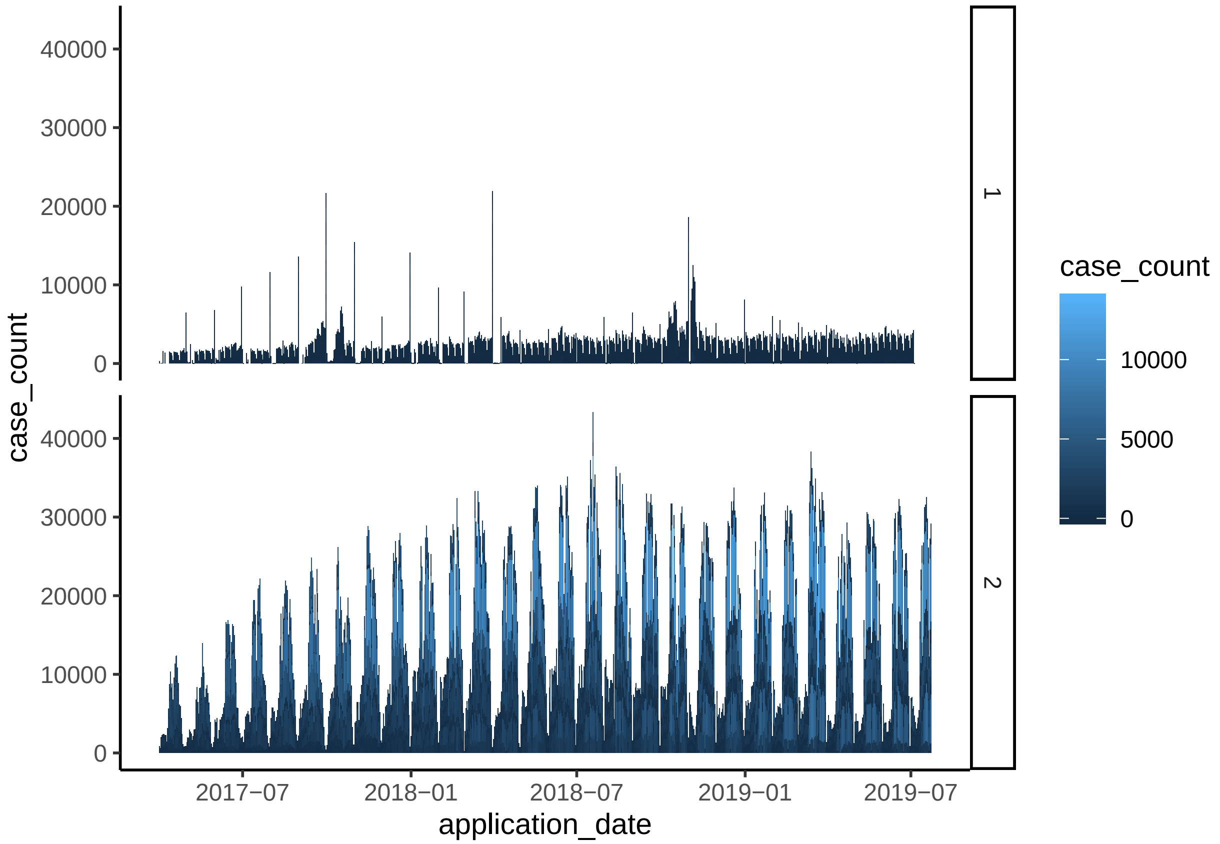
Visualization of data

Distribution of Business Segment



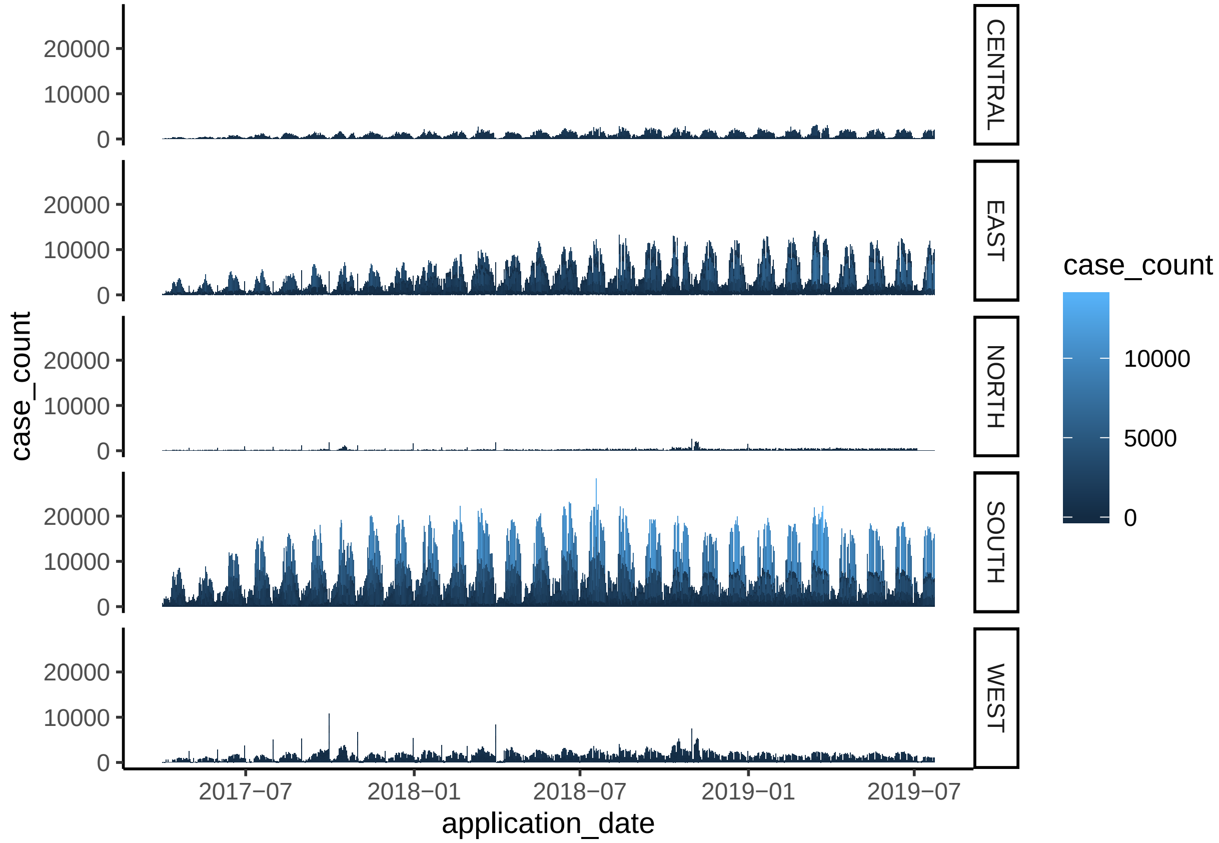
The Percent of Business Segment-1 is more than Business Segment-2

Number of cases as per Business Segment



As compared to Business Segment-1 there is more number of cases received in Business Segment-2 on daily basis.

Zone wise number of cases



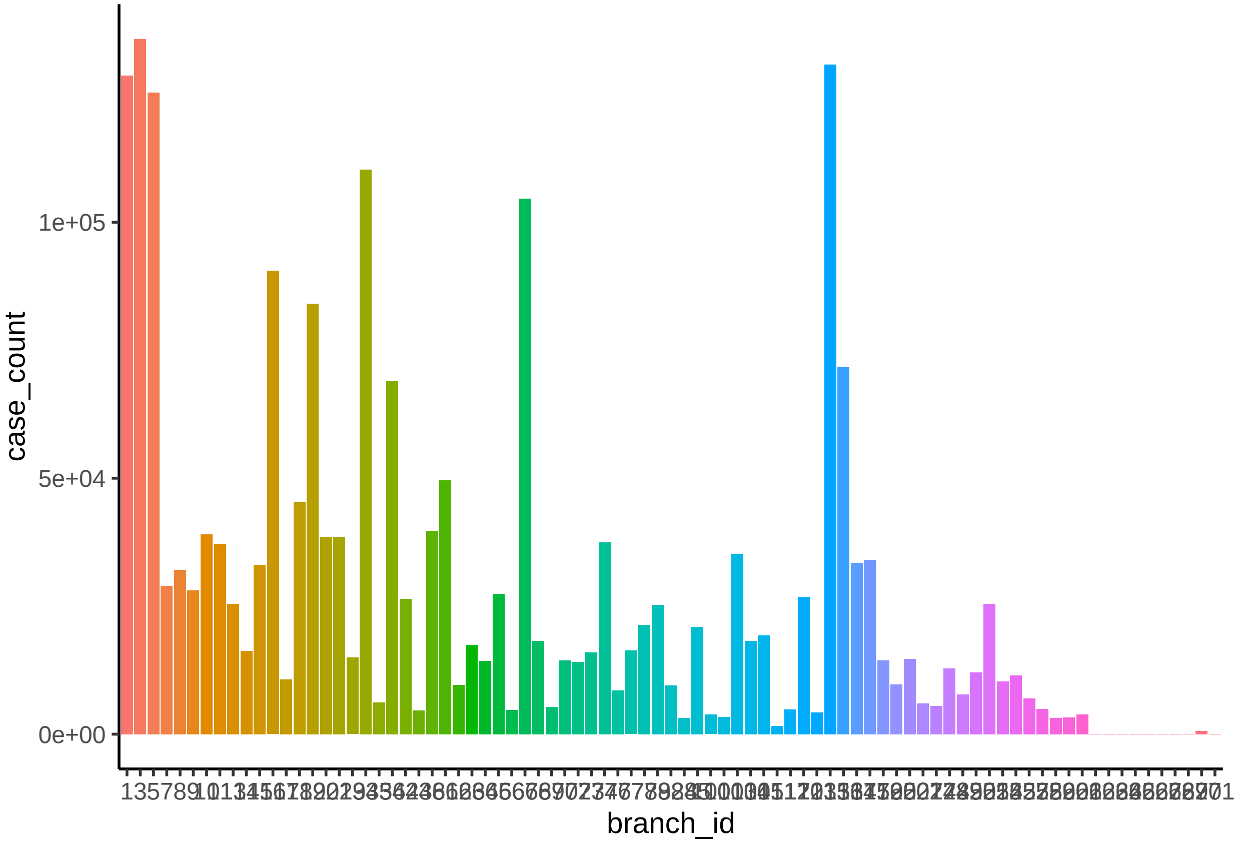
The Number of case_count is maximum in the SOUTH zone followed by EAST zone.

Number of branch_id in business segment 1

```
##
##   1   3   5   7   8   9  10  11  13  14  15  16  17  18  19  20  21  29  34  35
## 806 806 806 806 806 806 806 806 806 806 806 806 806 806 806 806 806 806 806 806
##  36  42  43  48  61  62  63  64  65  66  67  68  69  70  72  73  74  76  77  78
## 806 806 806 806 806 806 806 806 806 806 806 806 806 806 806 806 806 806 806 806
##  79  82  84  85 100 101 103 104 105 111 117 120 121 135 136 137 147 159 165 202
## 806 806 806 806 806 806 806 806 806 806 806 806 806 806 806 806 806 806 806 806
## 207 217 248 249 250 251 254 255 257 258 259 260 261 262 263 264 265 266 267 268
## 806 806 806 806 806 806 806 806 806 806 806 806 806 806 806 806 806 806 806 806
## 269 270 271
## 806 806 806
```

All branch id having equal count in Business Segment-1

Number of case_count as per branch_id



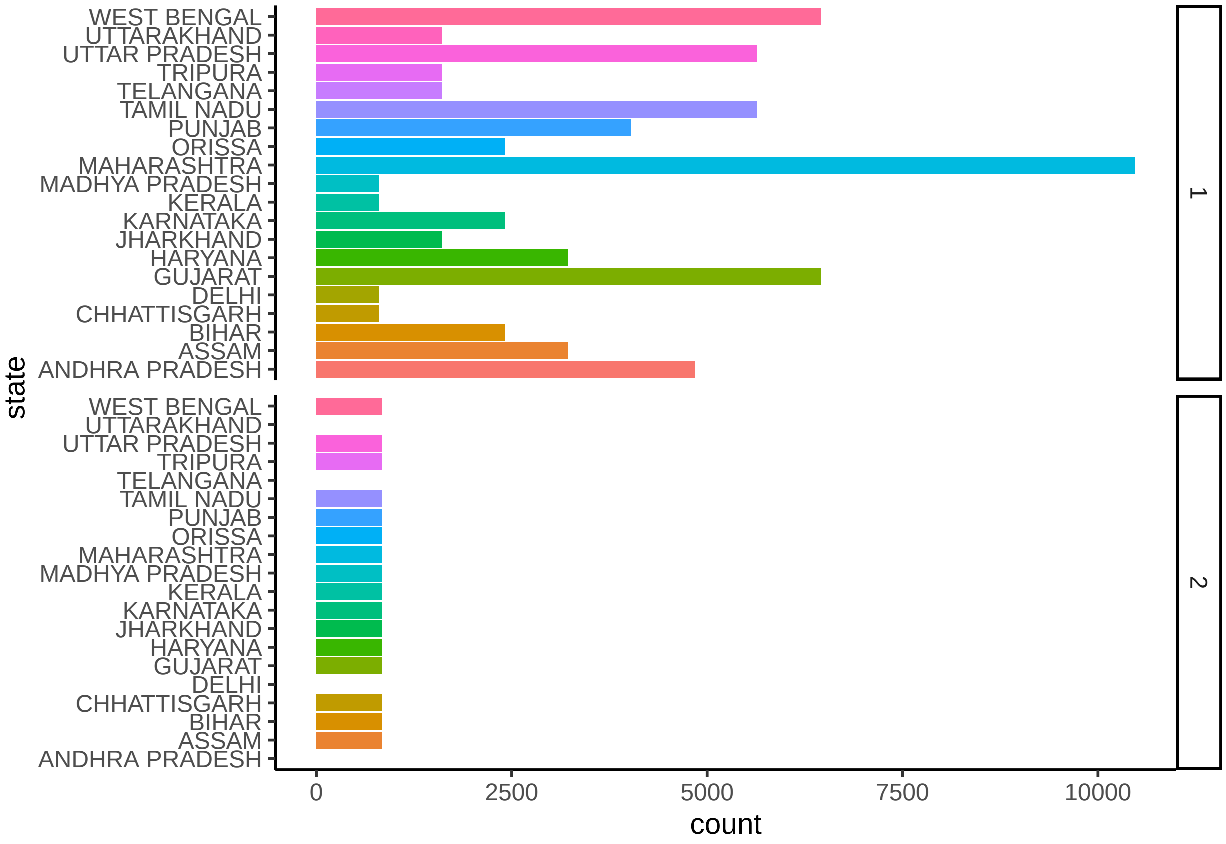
```
## [1] "branch_id with min number of case_count"
```

```
##    branch_id case_count
## 1:       263          0
## 2:       269          0
## 3:       267          1
## 4:       265          4
## 5:       262          5
## 6:       268          8
```

```
## [1] "branch_id with max number of case_count"
```

```
##    branch_id case_count
## 1:         3     135800
## 2:       135     130803
## 3:         1     128683
## 4:         5     125372
## 5:        34     110280
## 6:        67     104637
```

Number of state per business segment



In Business segment-1 more number of cases recieved but as we saw previously count of number of cases in Business Segment-2 is more than Business Segment-1.

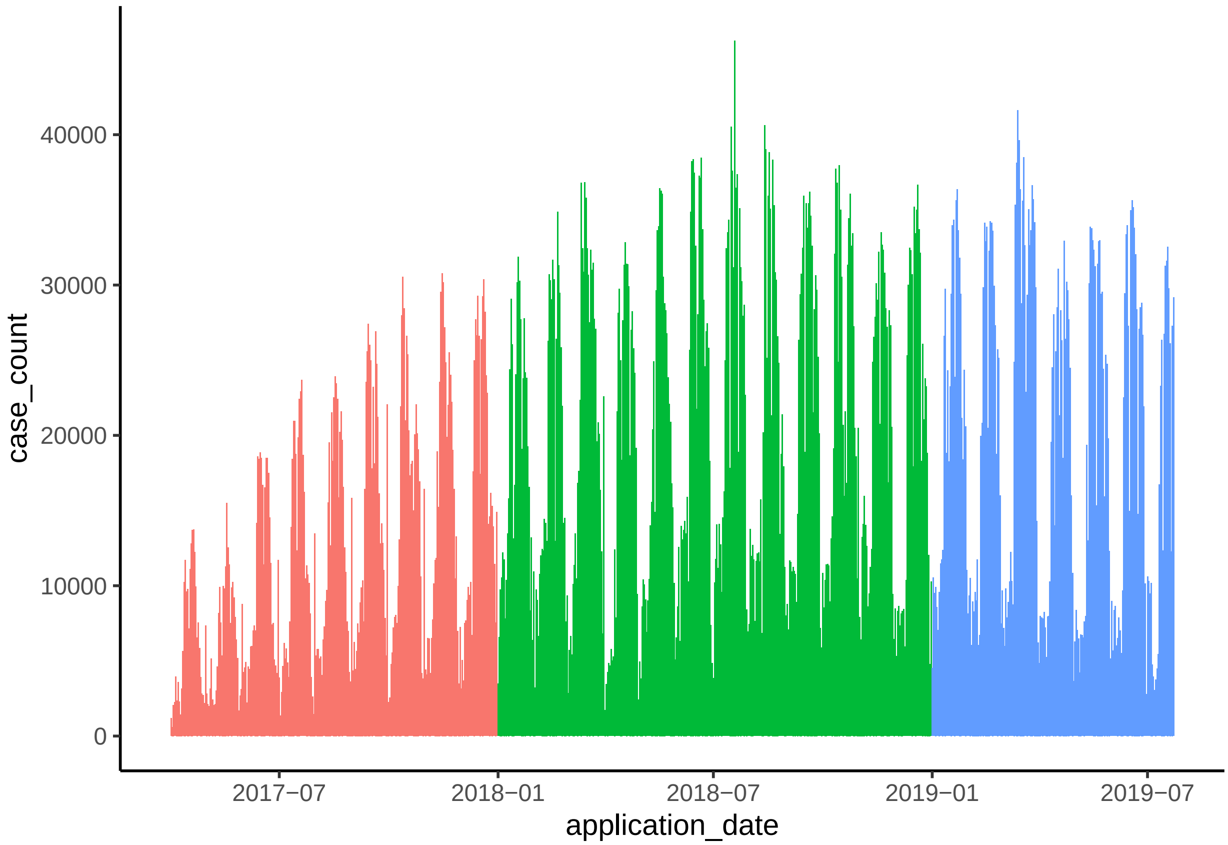
Feature Engineering

Steps involved in Feature Engineering:-

- Aggregate case_count by application_date
- Extract year,month, weekday and quarter from applicate_date
- Add Holiday Indicators to the data
- Collect Stock Information
- Collect daily India / U.S. Foreign Exchange Rate

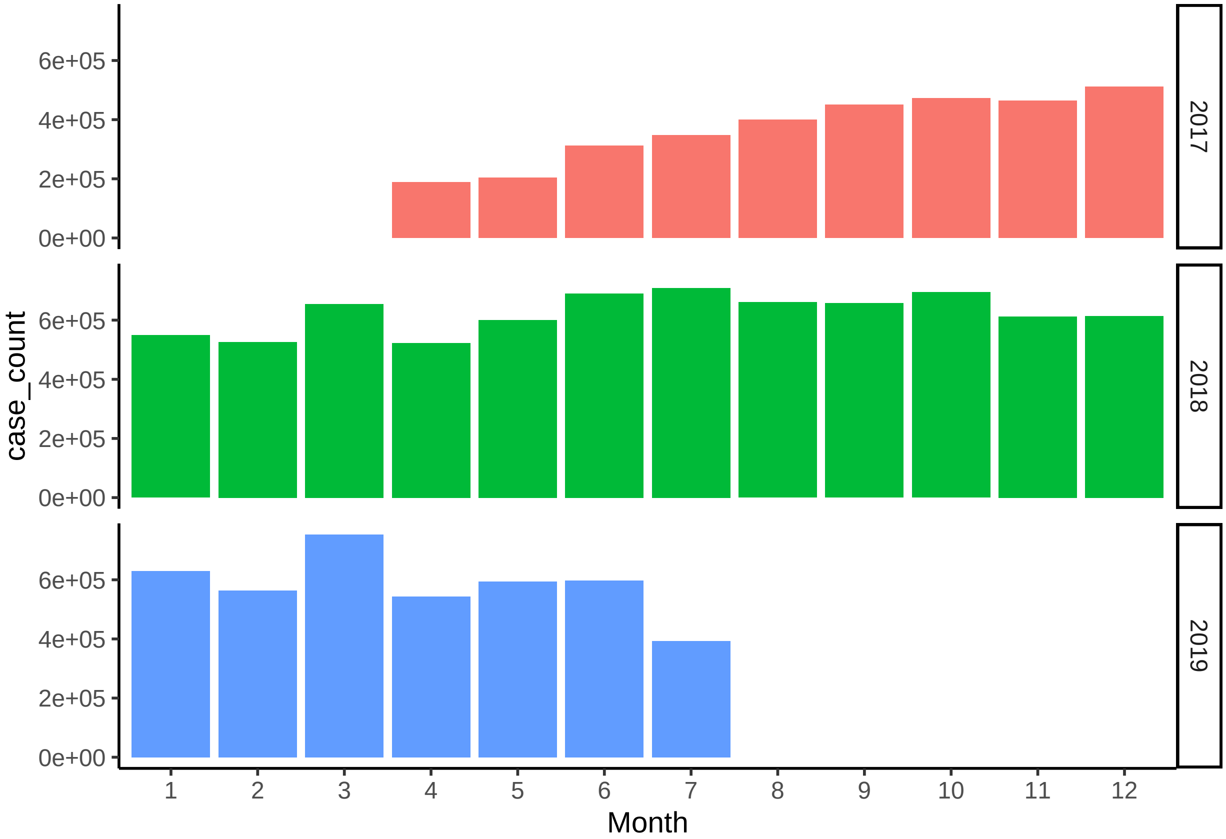
```
##    application_date segment case_count Year Month Day Weekday Quarter holiday
## 1:       2017-04-01       1        299 2017     4   1       7       2       0
## 2:       2017-04-03       1         42 2017     4   3       2       2       0
## 3:       2017-04-04       1         23 2017     4   4       3       2       0
## 4:       2017-04-05       1       1530 2017     4   5       4       2       1
## 5:       2017-04-07       1       1341 2017     4   7       6       2       0
## 6:       2017-04-12       1       1468 2017     4  12       4       2       0
##    LTFH.NS LTFH.BO    ER
## 1:    0.00    0.00  0.00
## 2:  113.15  113.70 65.10
## 3:    0.00    0.00 64.87
## 4:  115.80  115.65 64.58
## 5:  119.65  119.65 64.26
## 6:  120.20  120.15 64.55
```

Yearly Trend of Case_count



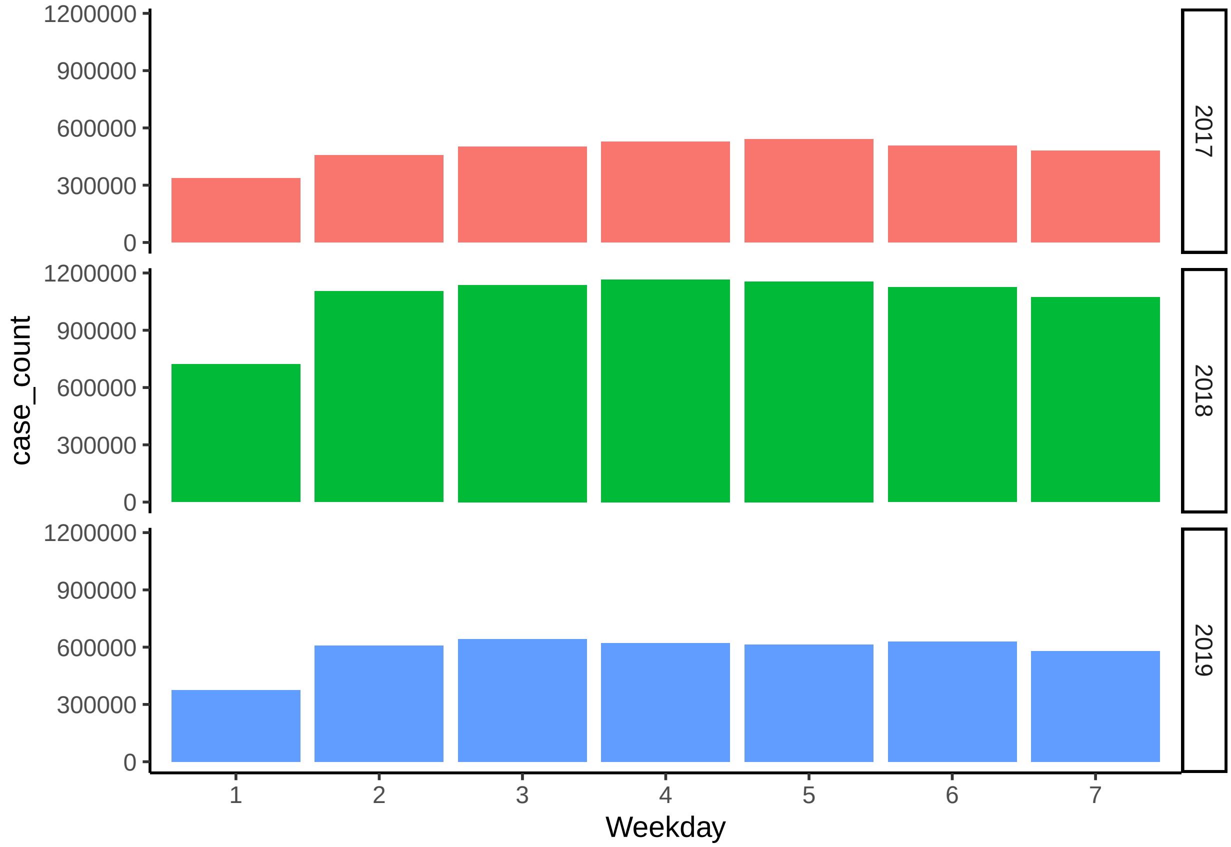
We have data starting from 2017-04 to 2019-07 and Number of case count showing slightly Positive trend with the year.

Monthly trend of Case_count



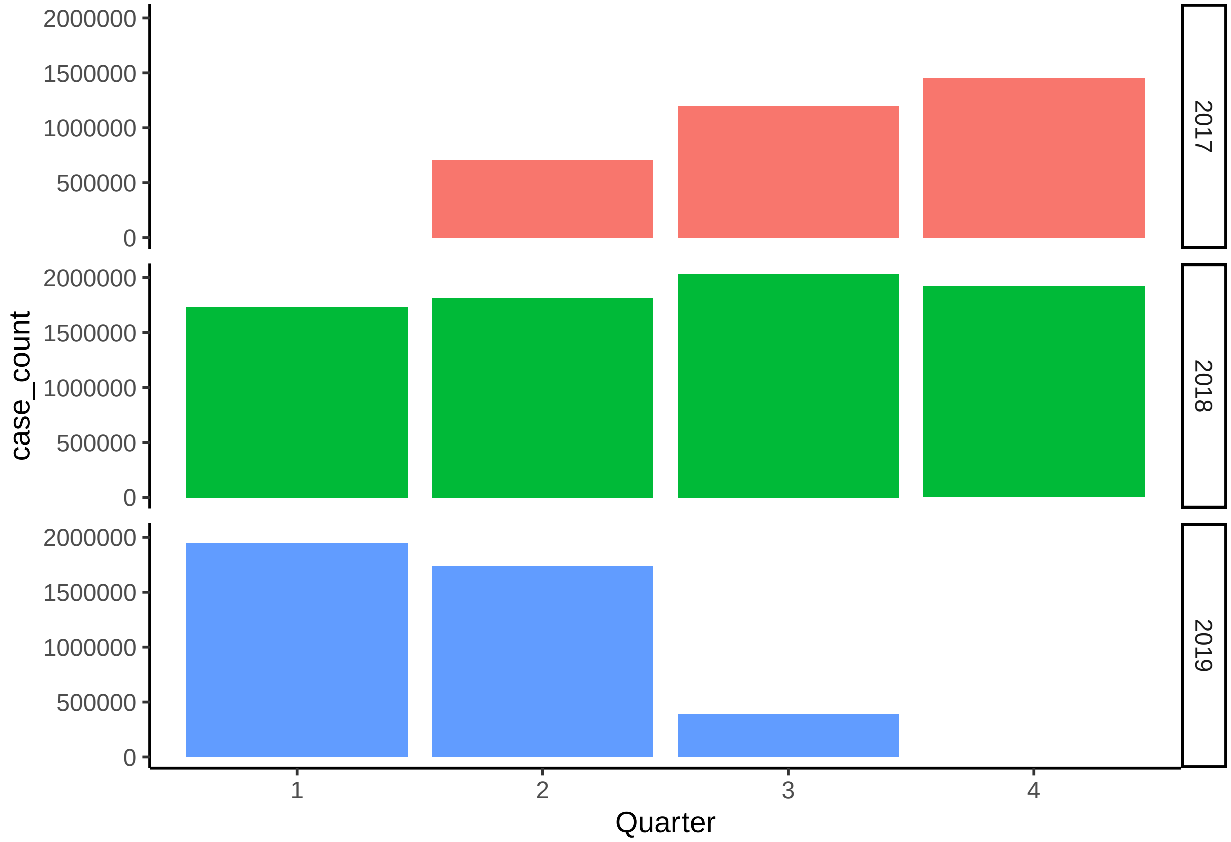
In the above graph, we can see that in 2017 the number of cases increasing with the month but in 2018 and 2017 showing ups and down in case_count.

Weekly trend of Case_count



Weekly Analysis of Number of case_count is Almost constant and on Weekday-1 (Sunday) there is vary less number of cases received.

Quarterly trend of Case_count



Here we can see that in 2017 and in 2018 there is a slightly positive trend in case_count.But in Q2 of 2019 there is decrease in number of cases.

Test data

```
##    id application_date segment
## 1:  1       2019-07-06       1
## 2:  2       2019-07-07       1
## 3:  3       2019-07-08       1
## 4:  4       2019-07-09       1
## 5:  5       2019-07-10       1
## 6:  6       2019-07-11       1
```

we need to Forecast for the dates provided in test set for each segment.

Feature Engineering of test data

Performed same data pre-processing as train data.

Model Building

For Model building we split the Data into train data and test data. We have 3 years of data so for daily forecasting we separate 2019 data as test dataset.

Predictive Model and Evaluation

- Regression with CNN Model

First:- Forecast for Business Segment-1

Extract the input dimension for the Keras model

```
## [1] 9 1
```

Model Fitting

```
## Model: "sequential"
## ________________________________________________________________________________
## Layer (type)                        Output Shape                    Param #
## ================================================================================
## conv1d (Conv1D)                     (None, 8, 64)                   192
## ________________________________________________________________________________
## conv1d_1 (Conv1D)                   (None, 7, 64)                   8256
## ________________________________________________________________________________
## conv1d_2 (Conv1D)                   (None, 6, 64)                   8256
## ________________________________________________________________________________
## conv1d_3 (Conv1D)                   (None, 5, 64)                   8256
## ________________________________________________________________________________
## flatten (Flatten)                   (None, 320)                     0
## ________________________________________________________________________________
## dense (Dense)                       (None, 64)                      20544
## ________________________________________________________________________________
## dense_1 (Dense)                     (None, 64)                      4160
## ________________________________________________________________________________
## dense_2 (Dense)                     (None, 16)                      1040
## ________________________________________________________________________________
## dense_3 (Dense)                     (None, 8)                       136
## ________________________________________________________________________________
## dense_4 (Dense)                     (None, 1)                       9
## ================================================================================
## Total params: 50,849
## Trainable params: 50,849
## Non-trainable params: 0
## ________________________________________________________________________________
```

```
##     loss
## 41725.26
```

Prediction on test data

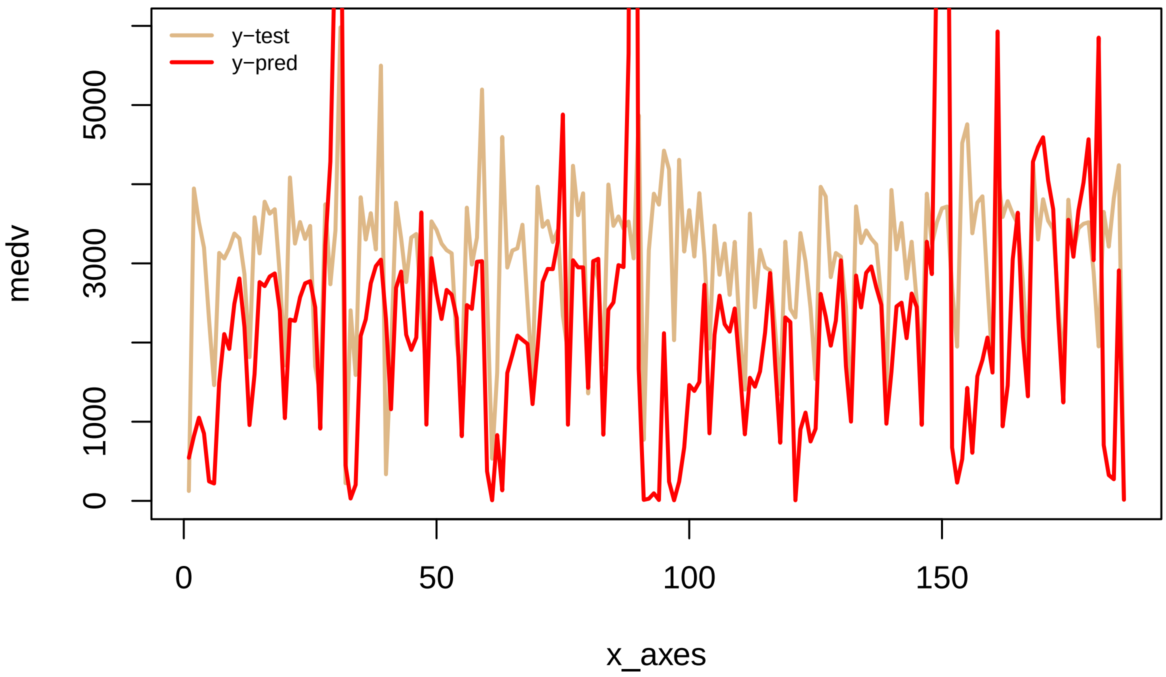
Evaluation Metric

The evaluation metric for scoring the forecasts is MAPE (Mean Absolute Percentage Error). The final score is calculated using MAPE for both the segments using the formula:

Final Score=0.5*MAPE(Segment-1)+0.5*MAPE(Segment-2)

```
## [1] 49.56896
```

Visualize Result



Second:-Forecast for Business Segment-2

Xgboost Model

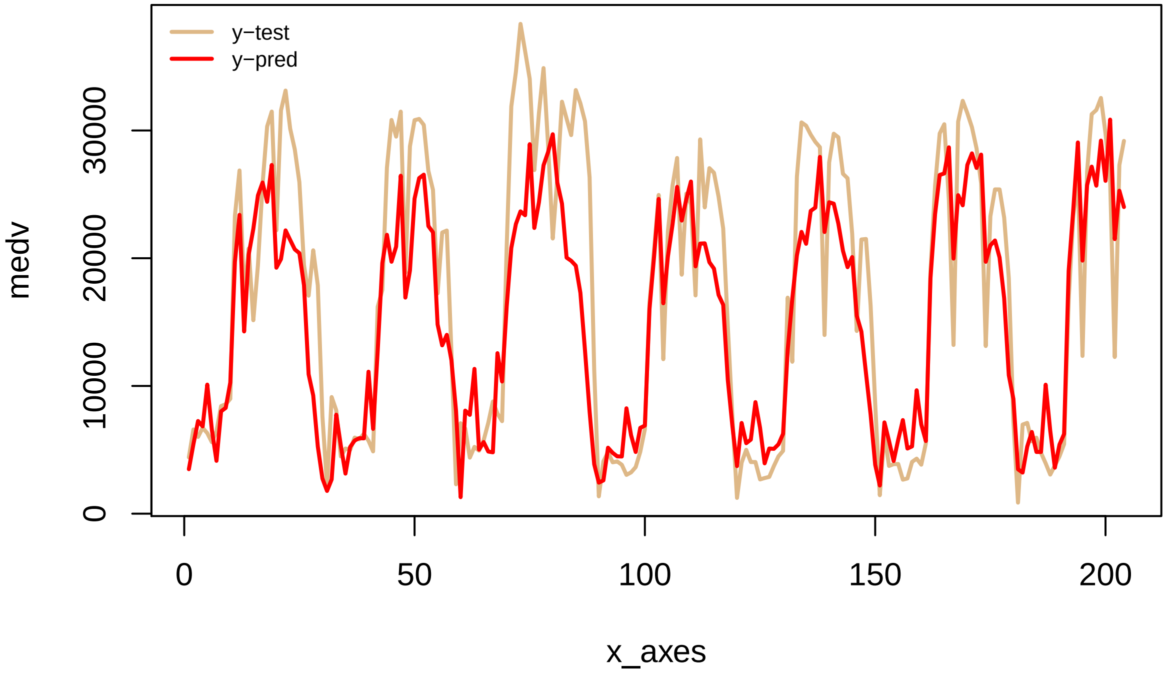
Prediction on test data

```
## [01:36:15] WARNING: amalgamation/../src/objective/regression_obj.cu:152: reg:linear is now deprecated
## [01:36:15] WARNING: amalgamation/../src/objective/regression_obj.cu:152: reg:linear is now deprecated
```

Evaluation

```
## [1] 34.22284
```

Visualize Result



Conclusion

In the end, we got our best model that can forecast daily cases for the next 3 months for 2 different business segments with MAPE of 49.56% and 34.22% respectively.