Capstone Project

Team 5

UCONN

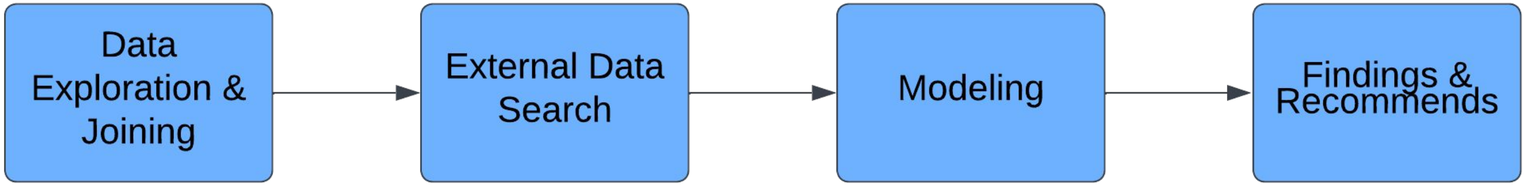
Introduction

Problem Description:

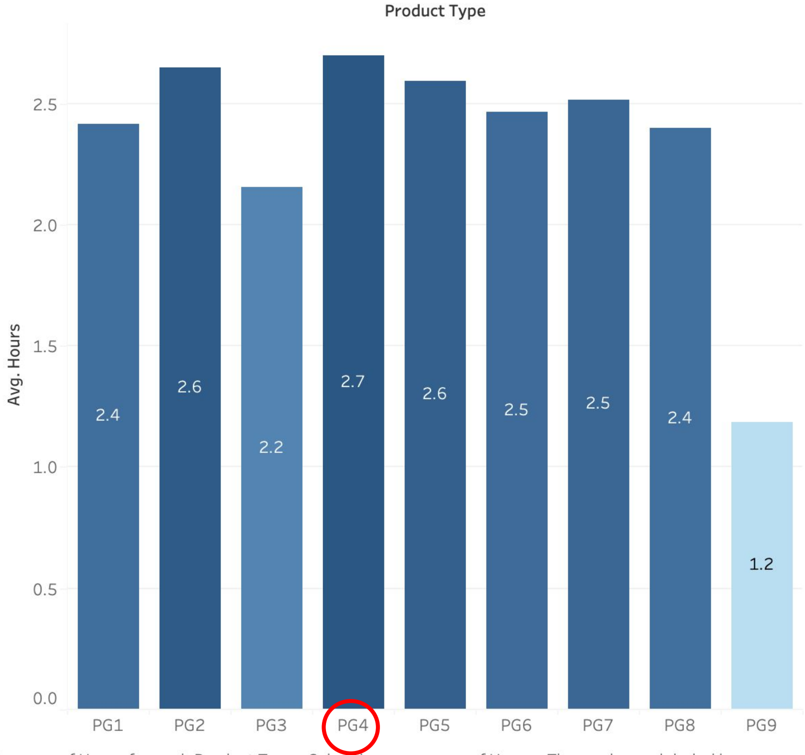
- Two types of maintenance: Scheduled and Callbacks.
- Callback disrupt daily operations due to high priority, random occurrence, and unknown repair time.

Objectives:

- To gain predictive insights into repair hours of callbacks,
- To provide data-driven predictions on weekly callbacks by office.

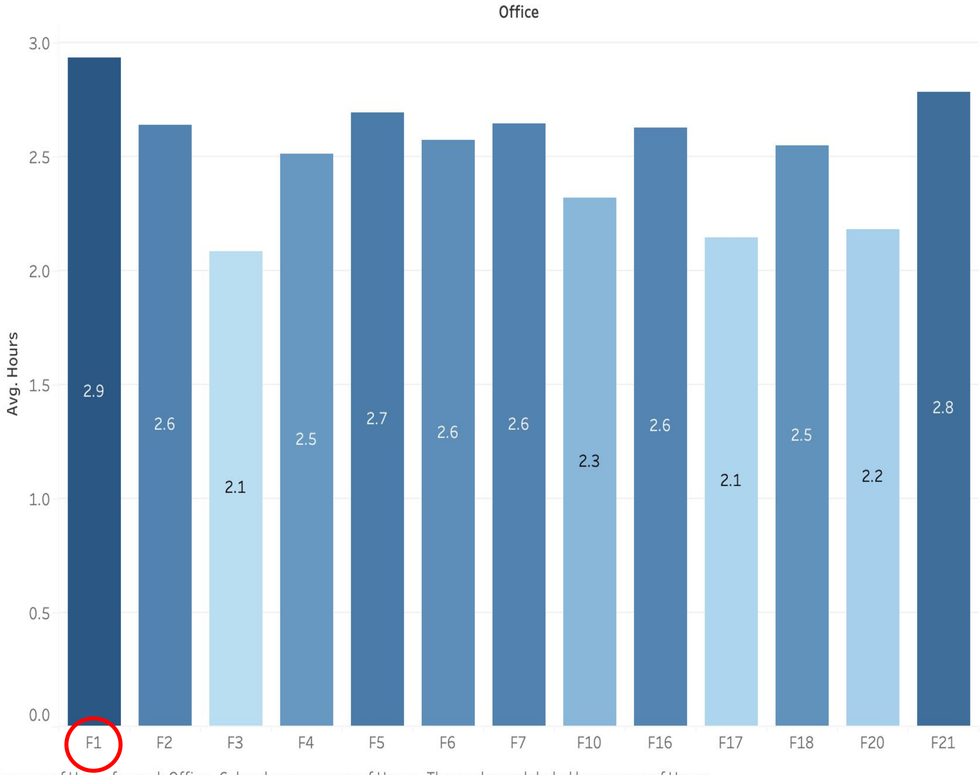


EDA - Most Hours



Average of Hours for each Product Type. Color shows average of Hours. The marks are labeled by average of Hours.

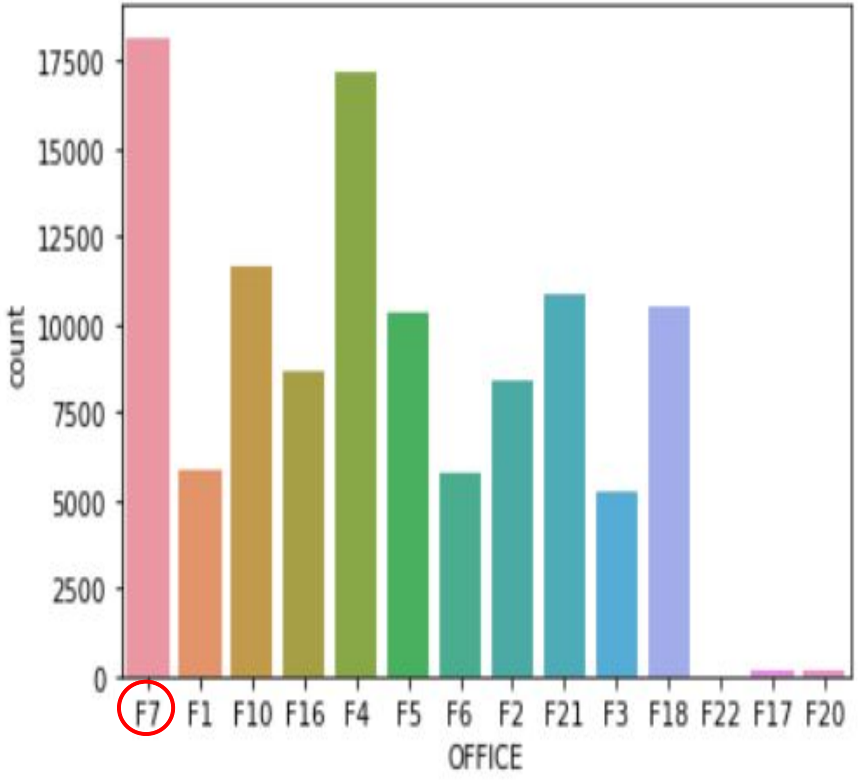
Product PG4 has the longest average hours



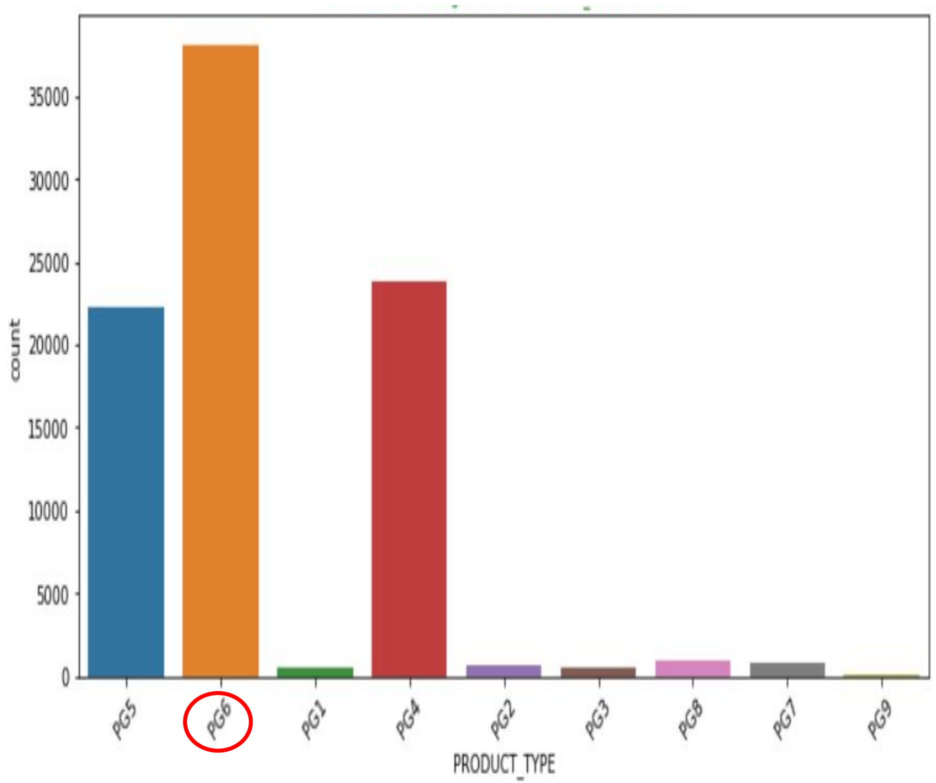
Average of Hours for each Office. Color shows average of Hours. The marks are labeled by average of Hours.

Office F1 has the longest average hours

EDA - Most Callbacks



Office F7 and F4 have the most callbacks



Product PG6 has the highest callback count

External Data

The external datasets used for further analysis and EDA on callback volume are:

1. **Population Data (US census data)**
   - Urban Population
   - Total Population

2. **Weather Data (https://open-meteo.com/)**
   - ✶ **longitude and latitude weather data mapped to select building zip codes and aggregated for the entire office**
   - Temperature
   - Relative Humidity
   - Precipitation Level

	OFFICE	ZIPCODE	LAT	LNG	CITY	STATE	time	temperature_2m (°C)	relativehumidity_2m (%)	precipitation (mm)
0	F1	02420	42.457055	-71.215464	LEXINGTON	Massachusetts	2021-01-03	0.609722	78.083333	0.181944
1	F1	02420	42.457055	-71.215464	LEXINGTON	Massachusetts	2021-01-10	-0.730357	63.994048	0.008929
2	F1	02420	42.457055	-71.215464	LEXINGTON	Massachusetts	2021-01-17	0.816071	74.696429	0.210714

GBO, Office, Building Type, Product type & City Majorly Impact Callback Hours

Best model - Gradient Boosting

- Joined the Time_ Tickets dataset with the Callbacks dataset on Work_date and Unit_ID - 62658 rows and 46 columns
- Our best model is Gradient Boosting
- Minimum Root Mean Square Error(1.566) and Mean Squared - Error(2.45) values

Top Features- Time taken to address a callback hours majorly depends on the following columns

1. **GBO**
2. **Office**
3. **Building Type**
4. **Product type**
5. **City**

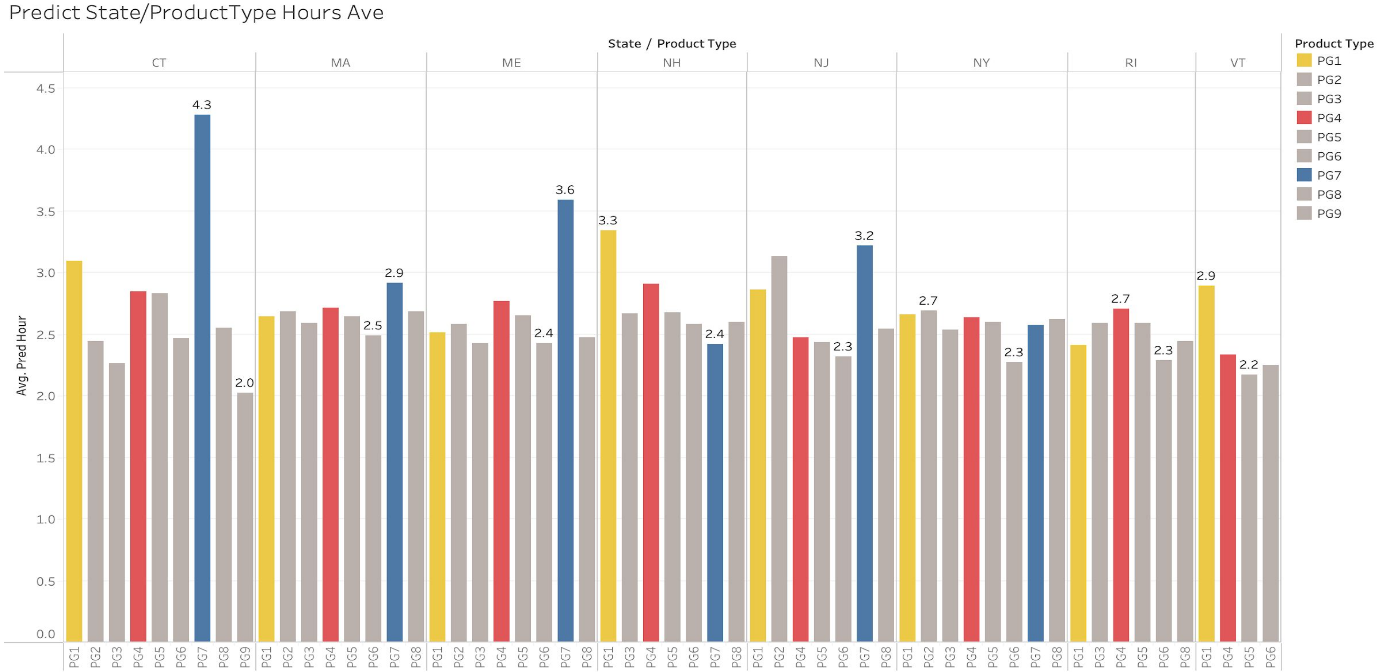
Prioritize PG7, PG1 and PG4

Observation:

- PG7, PG1 and PG4 have lowest count per state but highest time taken to fix the callback

Recommendations:

- Prioritize to PG7, PG1 and PG4. Include special training for mechanics if they lack the required skill sets or send more number of mechanics to fix the callbacks to reduce the repair time



Average of Pred Hour for each Product Type broken down by State. Color shows details about Product Type. The marks are labeled by average of Pred Hour.

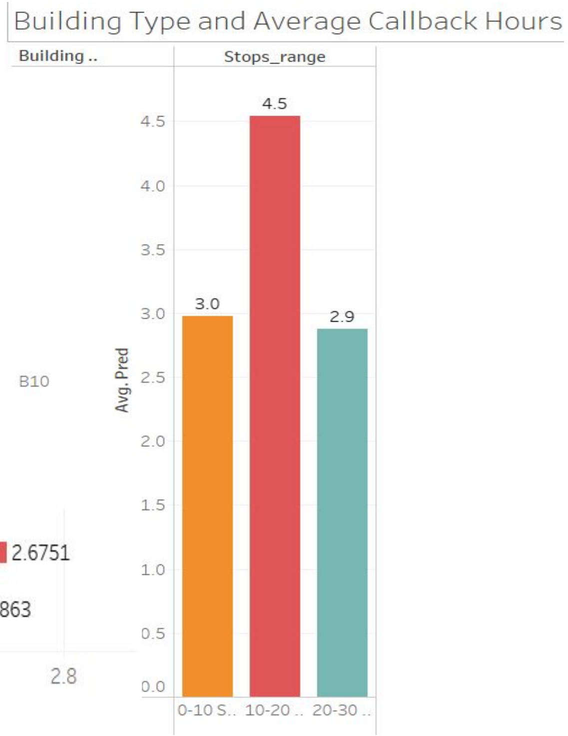
Building type B10 with 10-20 stops takes more time

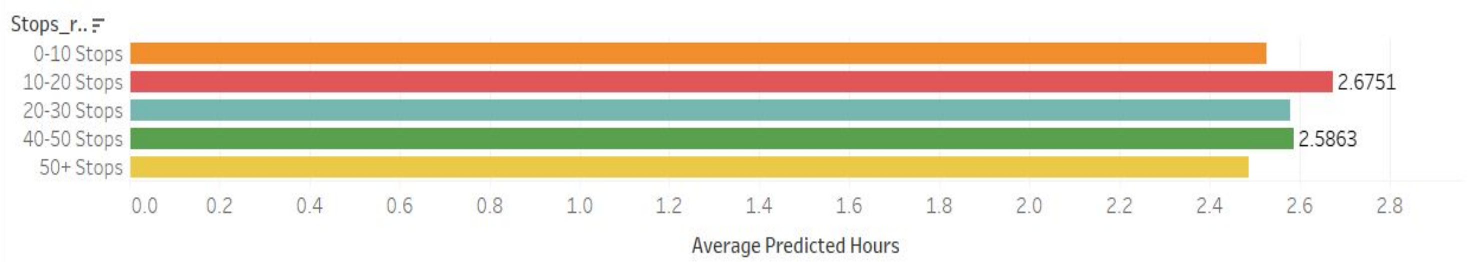
Observation:

- Buildings containing 10-20 stops taken more repair time, B10 takes highest repair time of 4.5 hours for 10-20 stops

Recommendations:

- Investigate the high repair time of callbacks with 10-20 stops and building type B10





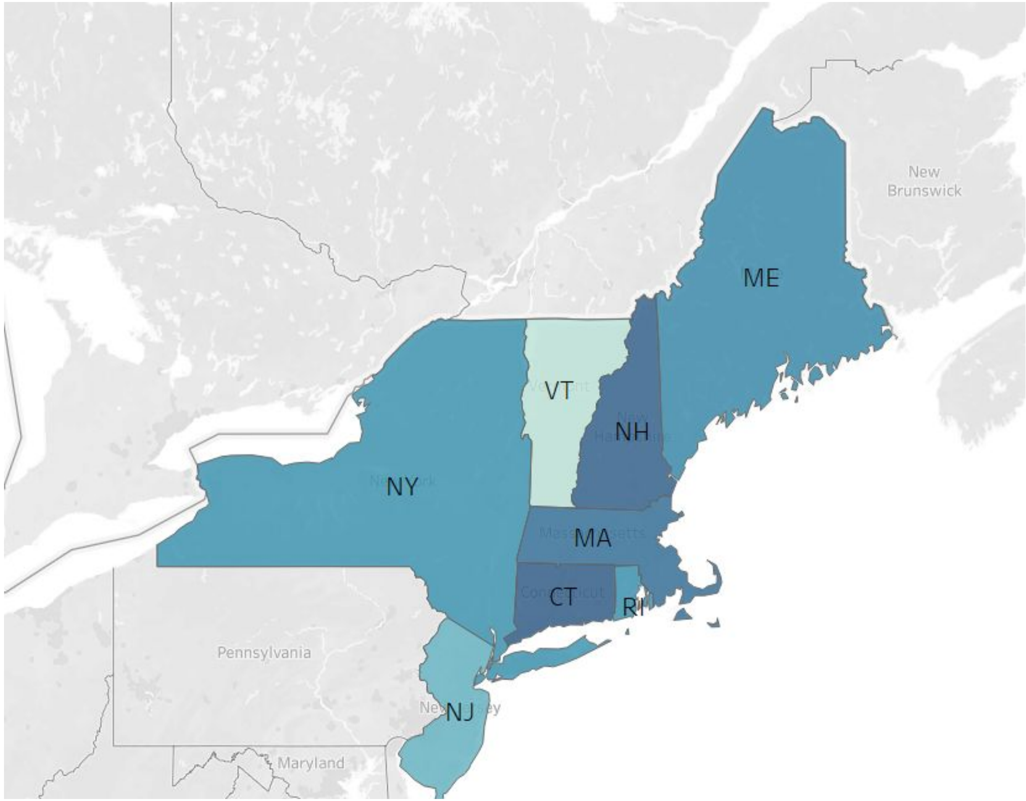
States NH, CT and MA take more callback hours

Observations:

- NH,CT,MA has high number of callback hours and high average callback hours compared to other states

Recommendations:

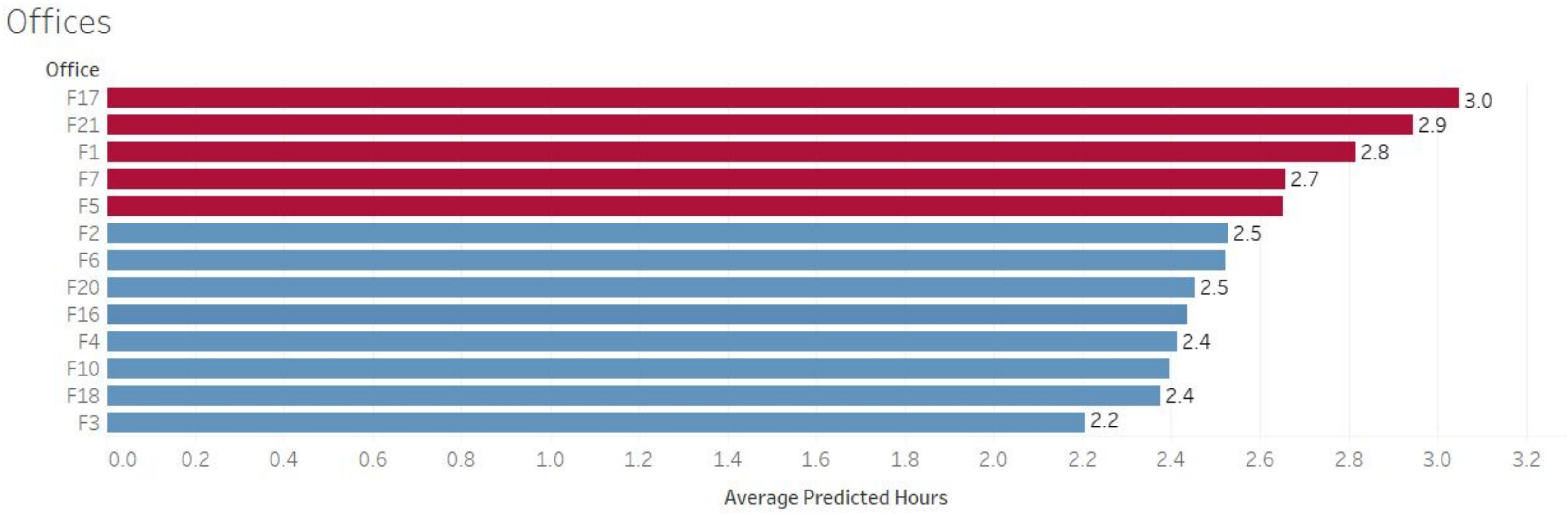
- We can have more mechanics in these regions so as to minimize the wait time and for handling high callback volume



Offices F7, F21, F5, F1 and F17 take more repair time

Observations:
- Offices F7, F21, F5, F1 and F17 take more than average time(2.5 hrs) to fix the callbacks

Recommendations:
- Assign more skilled mechanics to these offices to expedite the callback repair process



Predicting Weekly Callbacks - Methodology

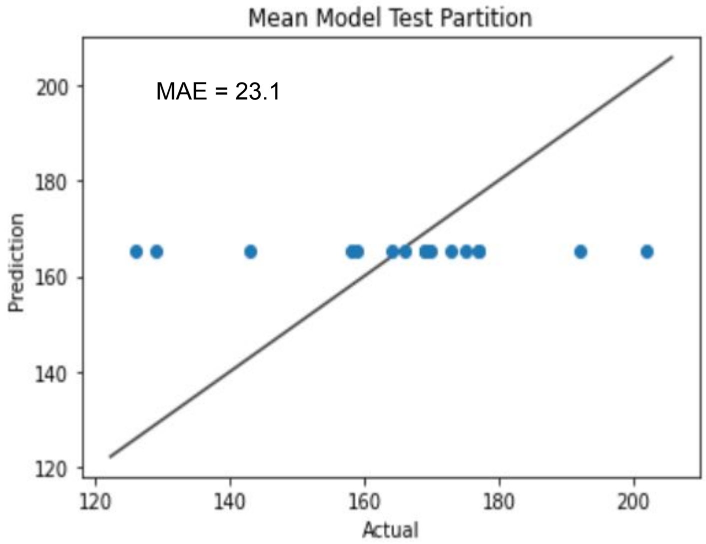
Our goal was to create a more accurate prediction for how many callbacks will occur at a given office for the next week:

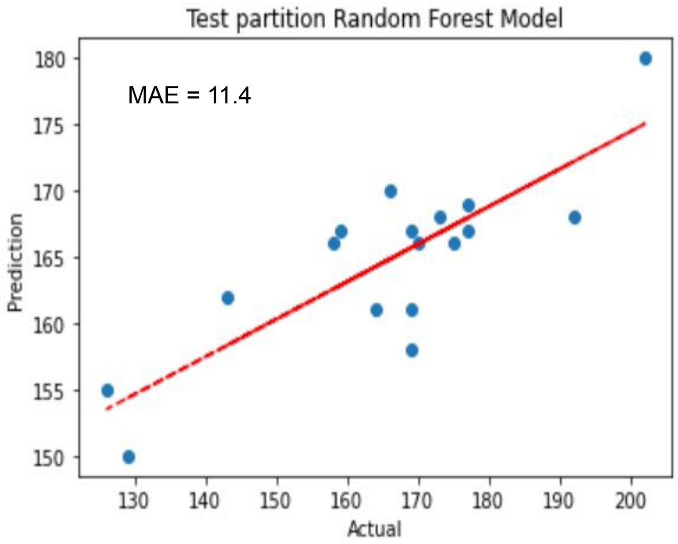
- Our first step was to create a baseline model
  - Mean only approach specific to each office
- Creating usable variables from the provided data and enhancing with external sources
  - Service proportion variable
  - Weather
  - US Census

$$\text{Service proportion: } \frac{\text{lag1(Sum of weekly service actions per office)}}{\text{(Total number of units in that office)}}$$

Predicting Weekly Callbacks - Office F4

- Our random forest model shows a clear lift compared to our baseline model
- Mean Absolute Square Error (MAE) is the average error in number of call backs
- Random errors



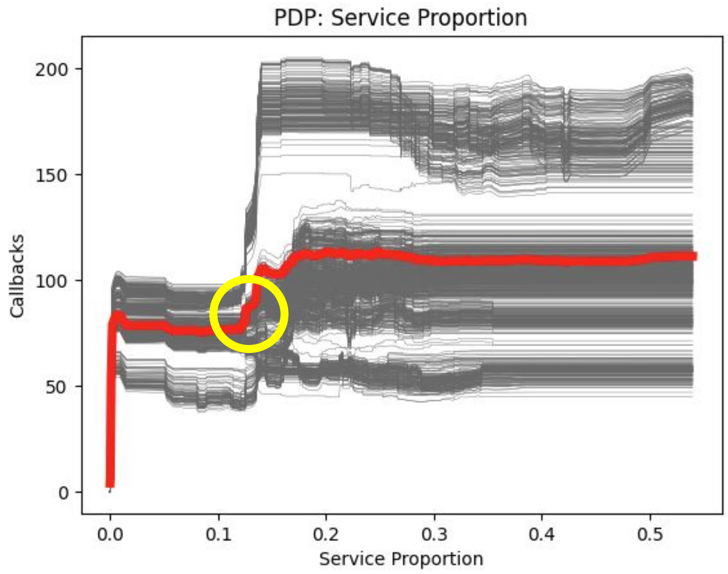


Predicting Weekly Callbacks - Results

- Average Mean Absolute Square Error 66% lower than baseline model
  - Only one small office was out performed by the mean only model
- Important Variables:
  - Service proportion - Significantly affected the model at lower values
  - Total Population - The model is affected by the largest population values
  - Avg Temperature (°C) - Lowest and highest temperature have slight affect callback volume

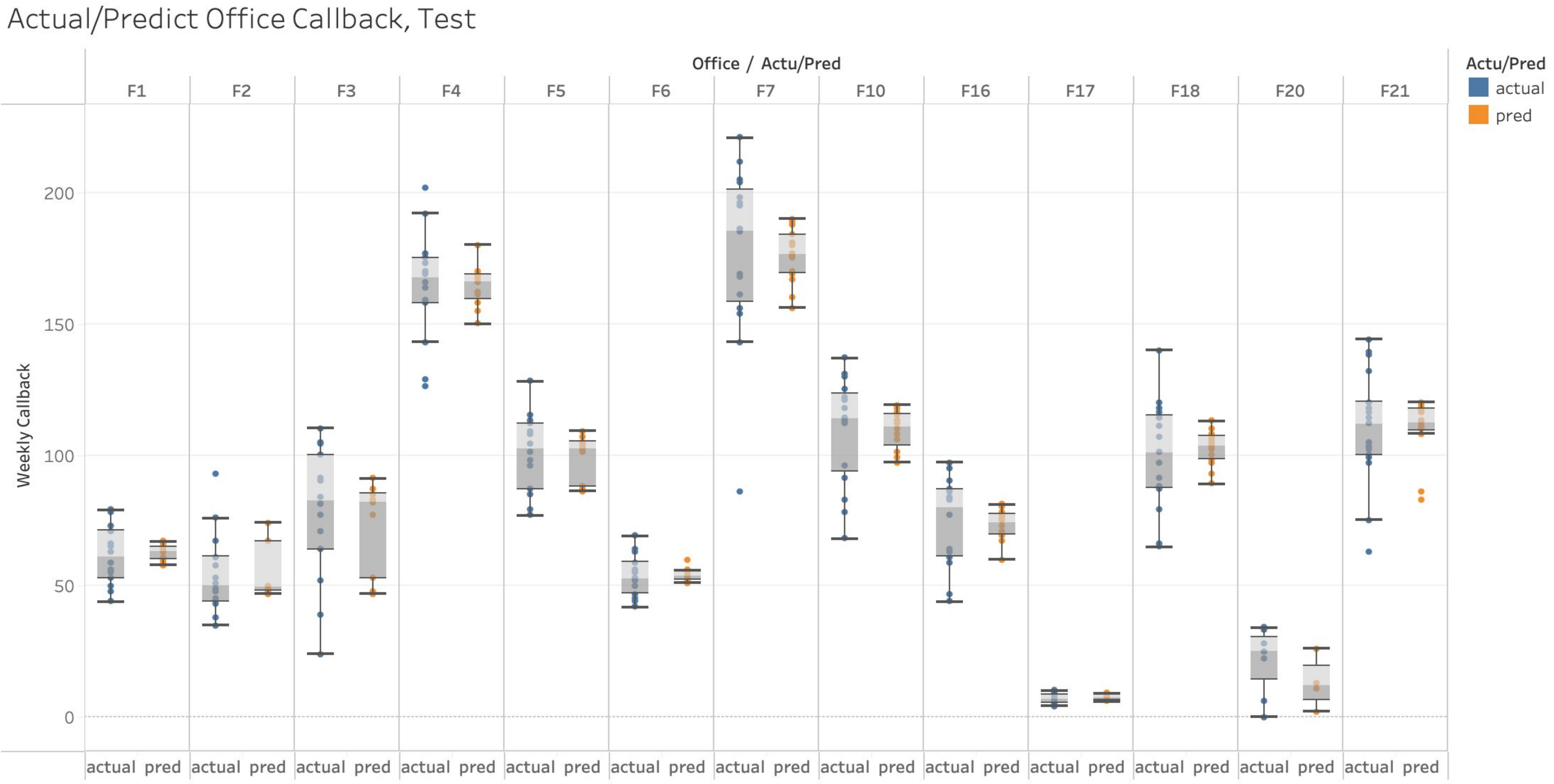
Recommendation Based on Partial Dependence Plot

- Based on our observations on average 10-12 percent of a office’s units were having scheduled maintenance a week
- The model starts to predict higher number of callbacks once the ratio exceeds this limit.
- This indicates that extra service actions are correlated with future callbacks, we recommend exploring units that were not subject to future callbacks and the difference between similar units that received the same type of repair but had a future callback



Service proportion: lag1(Sum of weekly service actions per office) / (Total number of units in that office)

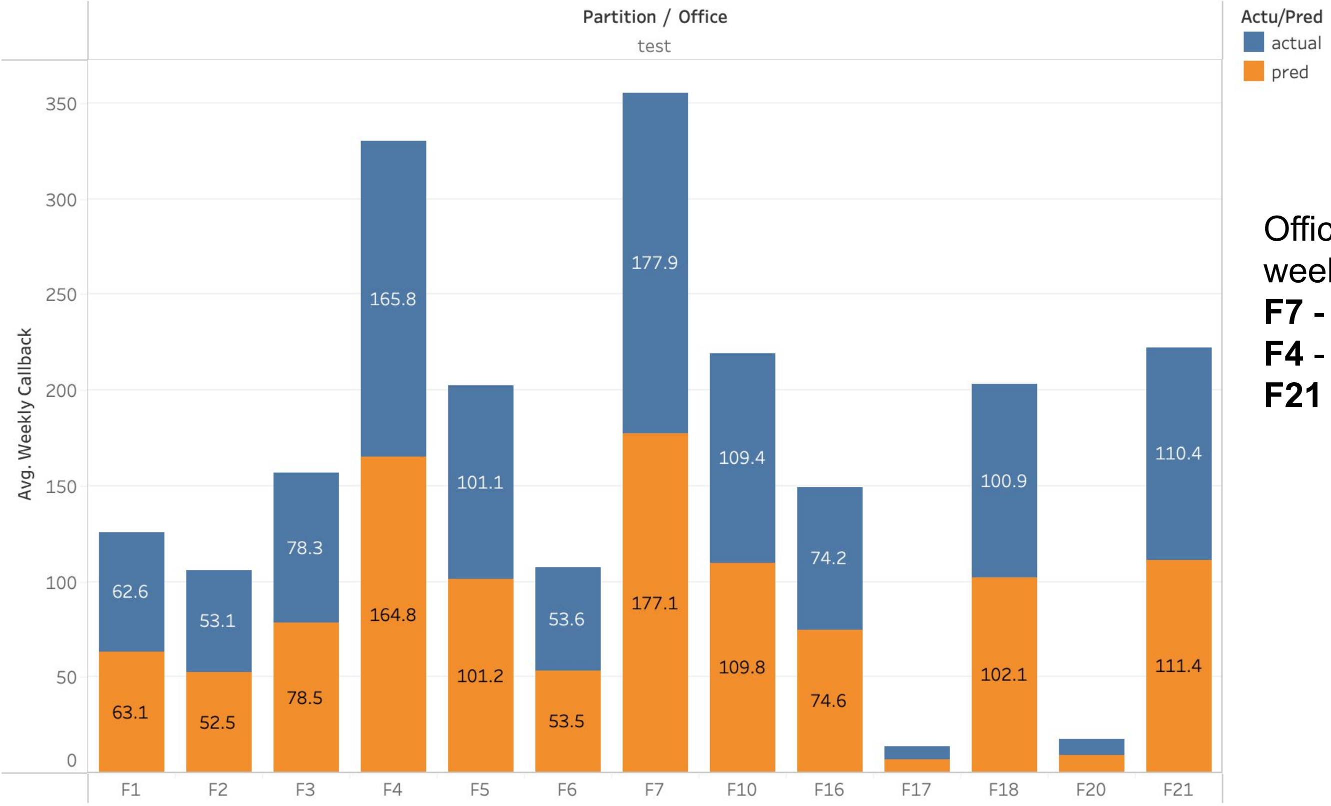
Predicted Weekly Office Callbacks



Weekly Callback for each Actu/Pred broken down by Office. Color shows details about Actu/Pred. The data is filtered on Partition, which keeps test.

Average Number of Weekly Callbacks Per Office

Average Weekly Office Callback Actual vs. Predict, Test Partition



Average of Weekly Callback for each Office broken down by Partition. Color shows details about Actu/Pred. The marks are labeled by average of Weekly Callback. The view is filtered on Partition, which keeps test.

Offices with highest count of predicted weekly callbacks:

F7 - 177.1

F4 - 164.8

F21 - 111.4

Questions or Comments?

Thank you!

Appendix

Data Joining

- Joined the Time Tickets dataset with the Callbacks dataset
- Casted Closed_DATETIME, Arrived_SiteTime, Callback_Date, DISPATCHED_DATE/TIME as date,
- Inner joined date with Work_Date from the Time Tickets dataset and UNIT_ID
- After joining have a total of 62658 rows and 46 columns

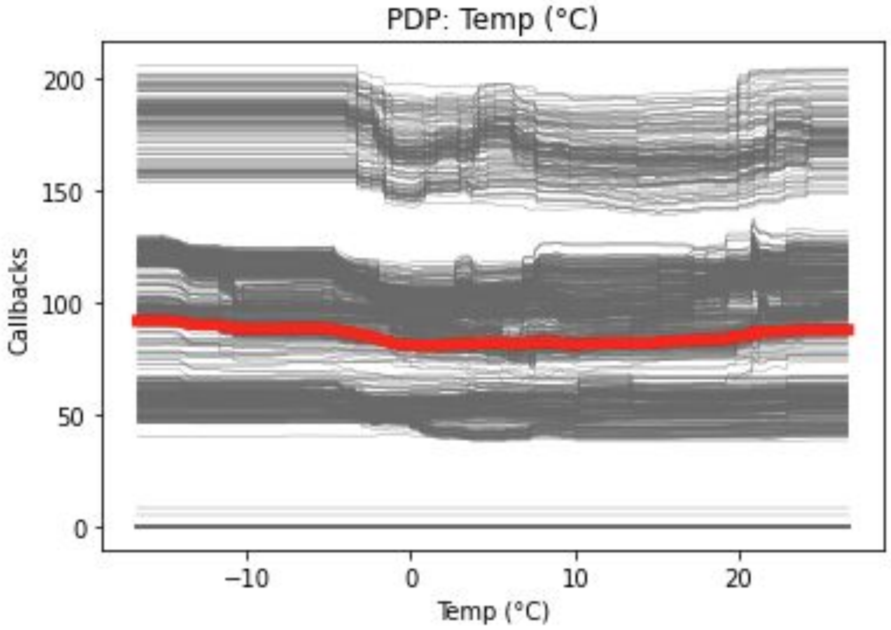
Callbacks Model Comparison - MAE

We're comparing Model Mean Absolute Square Error; this is the average difference between the number of predicted callbacks and the actual amount of call backs

The Random Forest Model outperformed the mean only baseline model across all offices except for office F20.

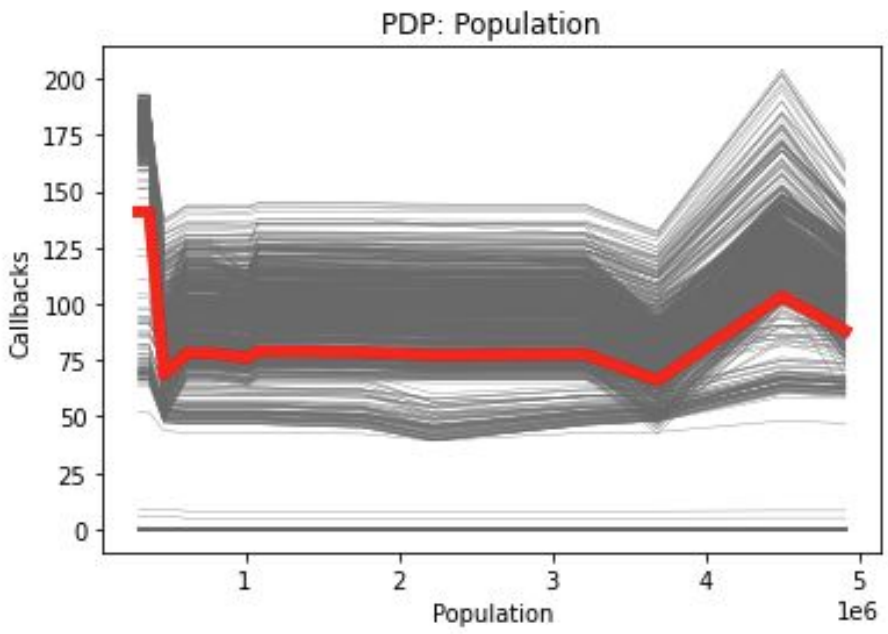
Partial Dependence Plot (Callbacks - Temperature)

Non-linear relationship between callbacks and temperature where the lowest and highest temperatures hav some influence on the model

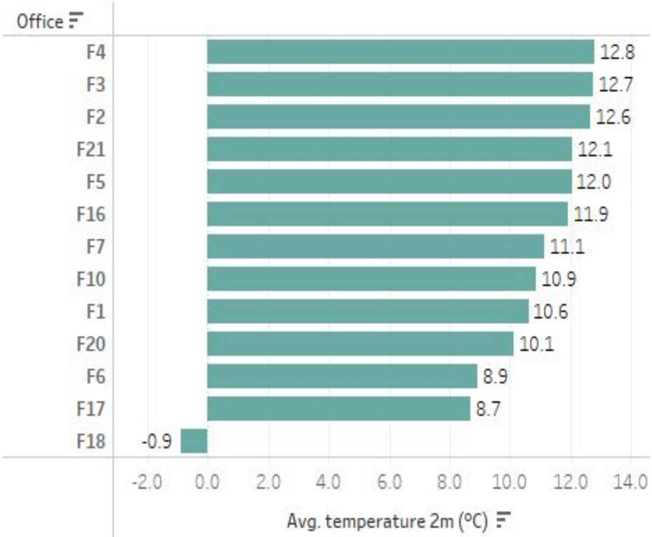


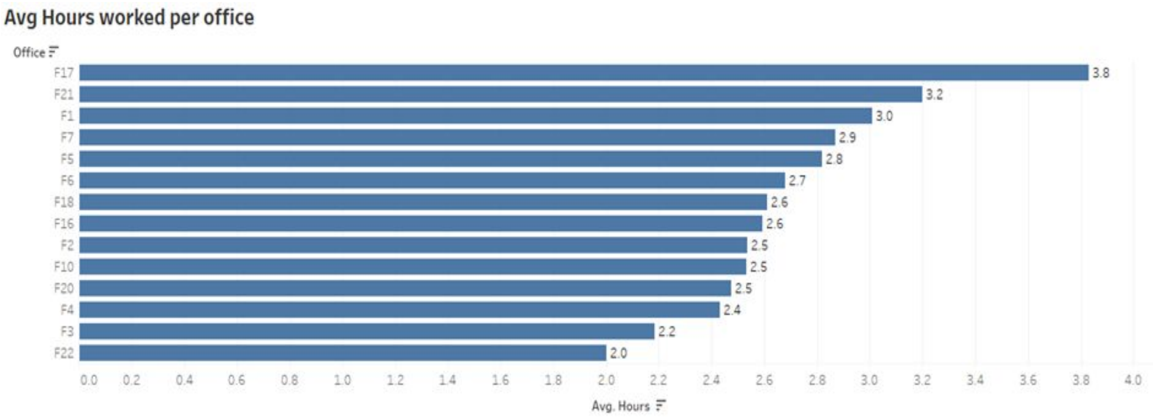
Partial Dependence Plot (Callbacks - Population)

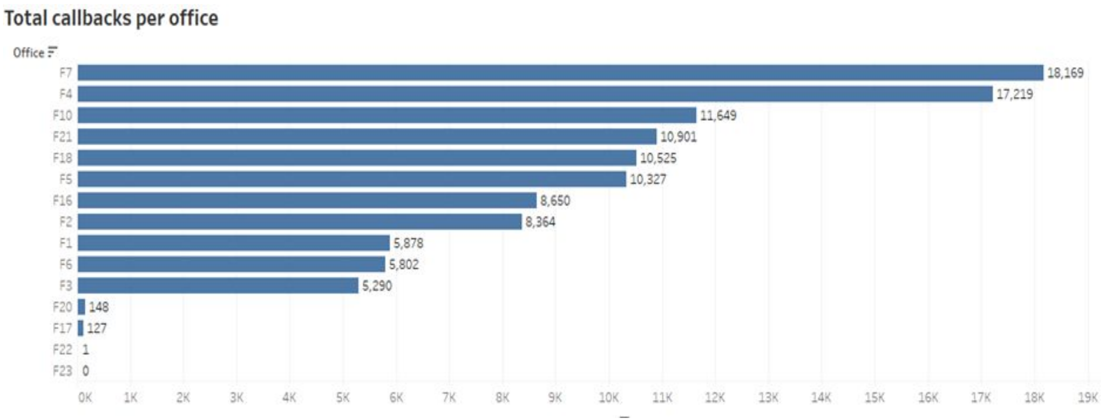
Non-linear relationship between callbacks and population where the highest populations have strong influence on the model.



External Data: Relationship Between Avg(temp) and Avg(hours) Per Office, Sum(callbacks) Per Office







1. Higher the temperature, lower is the average work hours worked at an office Eg - F3 & F4 , but this trend is not consistent.
2. Higher temperatures can also be related to higher callbacks per office. Eg - F4 & F7

Note : These patterns are not consistent.

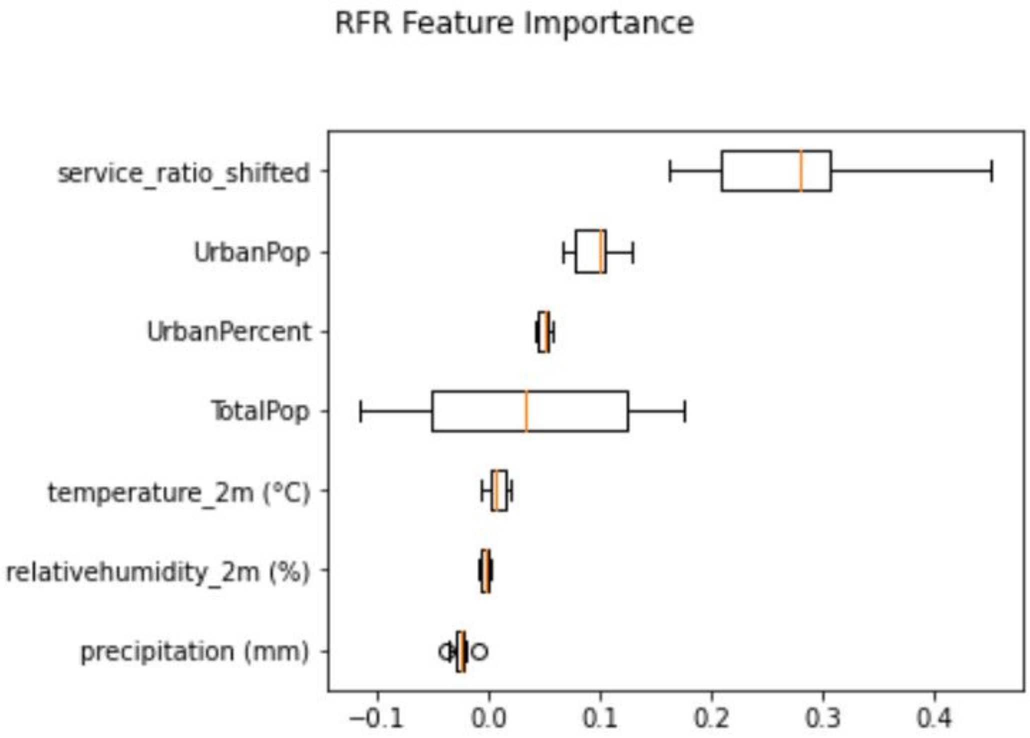
External Data: Relationship Between Callbacks and External Features

The size of the box plot denotes the impact of features like service_ratio_shifted, urban population, total population, temperature , relative humidity and precipitation on the no of callbacks.

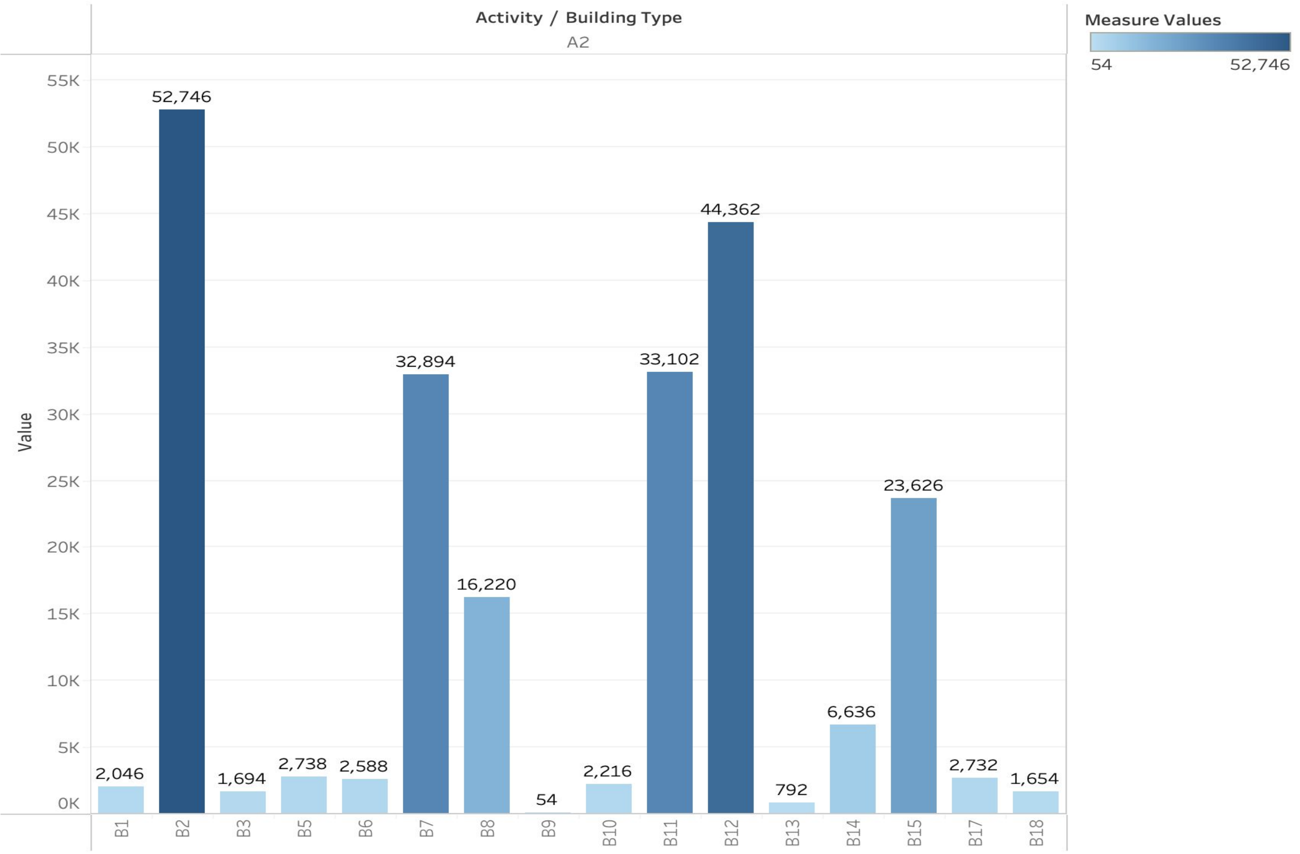
- Total population has a positive correlation with the no of callbacks indicating that offices in densely populated zip codes receive higher no. of callbacks.

- service _ratio has a considerable impact on no. of callbacks

- Temperature, relative humidity & precipitation have little or no impact on the no of callbacks.

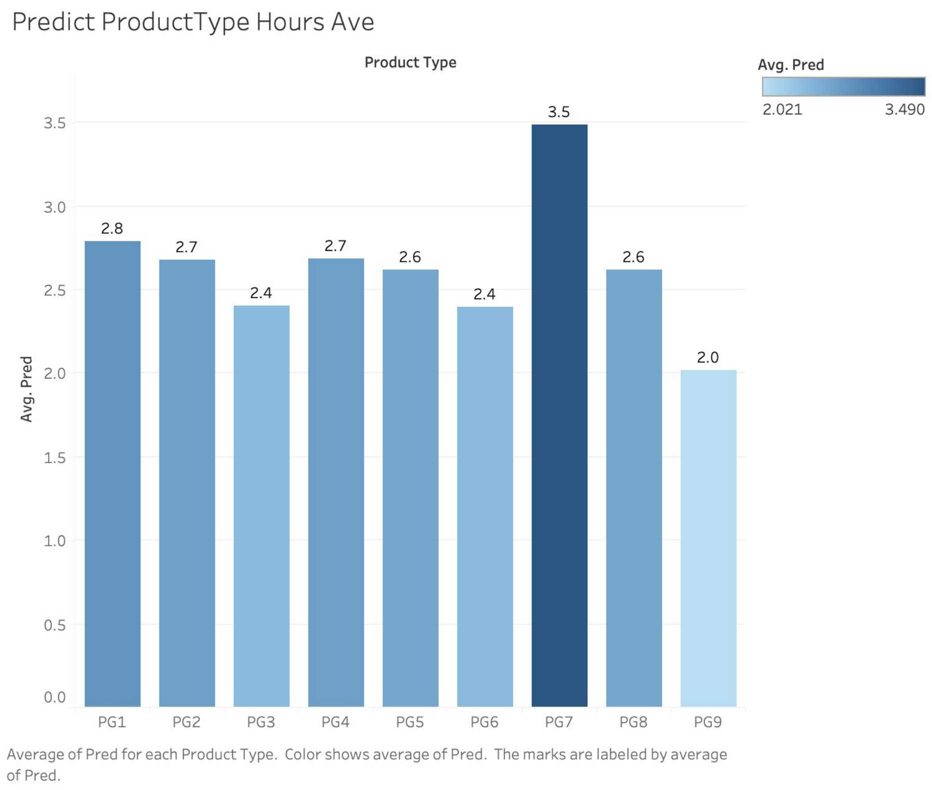


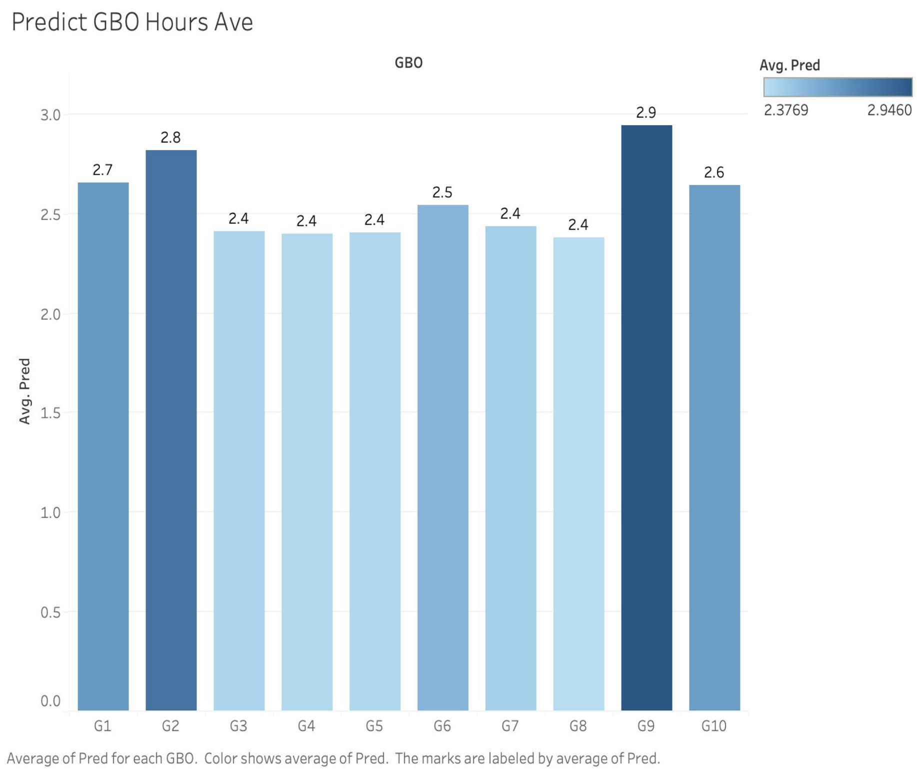
EDA - Most Callbacks



Count of Sheet1 for each Building Type broken down by Activity. Color shows count of Sheet1. The marks are labeled by count of Sheet1. The view is filtered on Activity, which keeps A2.

Predicted Hours Visualization





Model predicts Product Type PG7 has the longest hours

Model predicts GBO G9 has the longest hours