2020

Rodriguez, Jason W

WiFi Locationing - Predictive Modeling



# Purpose & Background

## Background

Many real world applications need to know the localization of a user in the world to provide their services. Therefore, automatic user localization has been a hot research topic in the last few years and continues to gain interest has technology becomes widely available. Automatic user localization consists of estimating the position of the user (*latitude, longitude, and altitude*) by using an electronic device, usually a mobile device. Outdoor localization problems can be solved very accurately due to the inclusion of GPS sensors into mobile devices. However, indoor localization is still an open problem mainly due to the loss of GPS signal within the indoor environment. Although, there are some indoor positioning technologies and methodologies, the dataset used is focused on WLAN finger-based ones (*aka WiFi Fingerprinting*).

## Business Challenge

A new IOT Analytics customer is developing a new system to be deployed to a variety of large campuses; such as shopping malls, universities, industrial campuses, etc… that will help users navigate the complex and unfamiliar interior spaces without getting lost. The challenge is that existing GPS technology works reliably well outdoors but generally does not work once the users are indoors. A different technology approach is being explored that investigates the feasibility of leveraging “WiFi fingerprinting” to determine a person’s location within interior spaces of a building.

The business challenge that IOT Analytics needs to explore is defining and evaluating multiple machine learning models to identify which produces the best results. Once we identify the best performing model, we will recommend that machine learning model to our customer so it can be incorporated into their development and deployment of the new app technology.

# Goals / Objectives

The following describes the goal(s) and objective(s) the team set out as part of the project charter.

## Goals

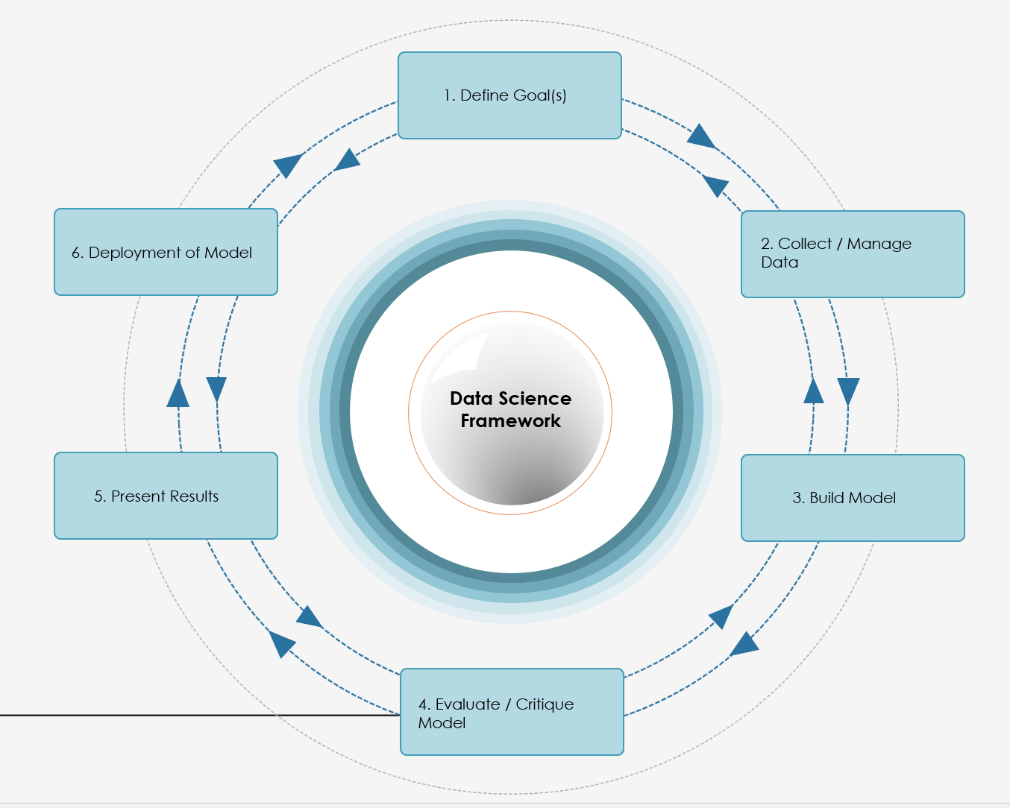
* Explore three (3) different machine learning models using the existing dataset provided by the client
* Identify the best performing models and make recommendations
* Best model should achieve performance metrics of 80% or better

## Objectives

* Apply a standardize data science methodology / process that provides transparency into the framework leveraged
* Provide the customer with all results along with supporting code so future evaluations and enhancements can be made

# Data Science Process

The following data science process framework was applied into our project execution. This execution framework is an iterative approach of moving through the various phases in order to acquire, process, define, and evaluate machine learning models. This methodical approach navigates the team to identifying the best model to meet our goals and objectives while provide guidance toward our execution. It should be noted this framework is the “Practical Data Science with R” framework as described by Zumel and Mount.

1. ***Define Goal(s)***

* *What are we solving?*
* *What limitation / constraints exist?*
* *SMART Goals*
* *What solutions is being requested?*

1. ***Collect/Manage Data***

* *Data Acquisition*
* *Pro-Processing*
* *Exploratory Data Analysis*

1. ***Build Model(s)***

* *Model Definition*
* *Model Development*

1. ***Evaluate/Critique Models***

* *Model Examination*
* *State gate decisions*

1. ***Present Results***

* *Report Development*
* *Stakeholder Communication and Analysis*

*Figure 1*

1. ***Deployment of Model***

* *Deployment Strategies & requirements*
* *Translation to deployment teams*
* *Model documented*

# Data Overview

## Dataset

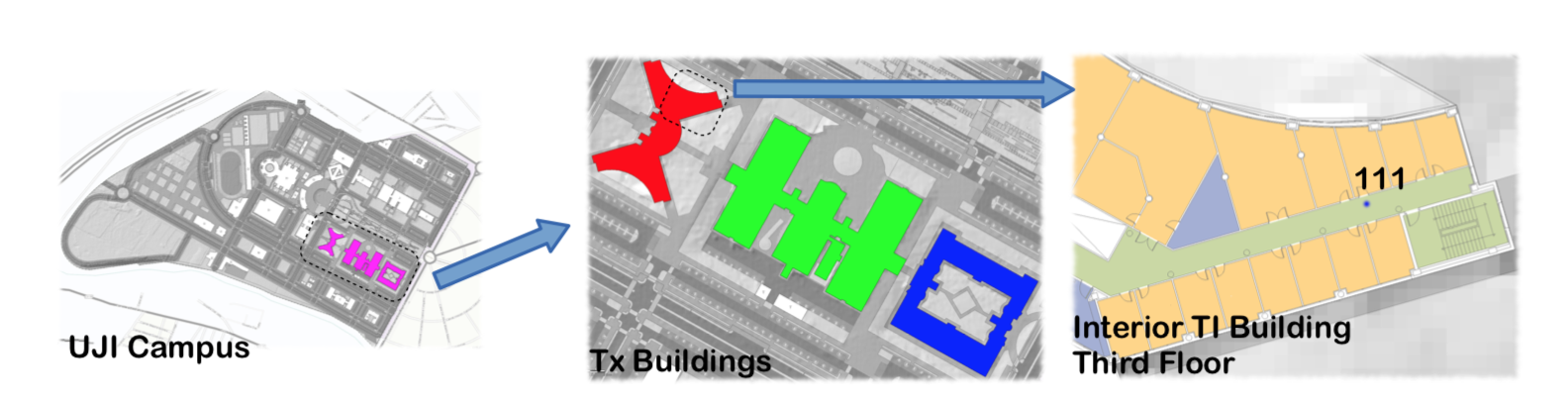
The original dataset provided by the customer is ***UJIIndoor Loc Dataset***. The UJIIndoorLoc dataset encompasses three different buildings at Universitat Jaume, which each building includes different spatial layouts, different number of floors, and approximately 108,700m2 of floor space combined. The overall database consists of **19,937** training/reference records (*trainingData.csv*) and **1,111** validation/test results (*validationData.csv*). The data was originally created in 2013 by means of more than 20 different users and 25 Android devices. 529 attributes contain the WiFi fingerprint, the coordinates where the signal was captured, and other useful information identifying the WiFi fingerprint.

Each WiFi fingerprint can be characterized by the detected Wireless Access Point (WAPs) and the corresponding Received Signal Strength Intensity (RSSI). The intensity values are represented as negative values ranging from -104dBm (*extremely poor signal*) to 0dBm (*extremely strong signal*). The positive value 100 is used to denote when a WAP was not detected. The database contains 520 different WAPs, thus the WiFi fingerprint is composed by 520 intensity values.

Source: <http://archieve.ics.uci.edu/ml/datasets/UJIIndoorLoc>



*Figure 2*

The more important features for WiFi fingerprinting are the WAPs detected and their RSSI intensity level value. ‘SpaceID’ contains a single integer value, in this case, it is used to identify the particular space (*offices, labs, etc..*.) where the signal capture was taken. ‘Relative Position’ indicates where the signal was taken to determine if the signal capture was inside the room (*represented by a value of 1*) or outside the space/room in the hallways (*represented by a value of 2*), as seen in Figure 3 below

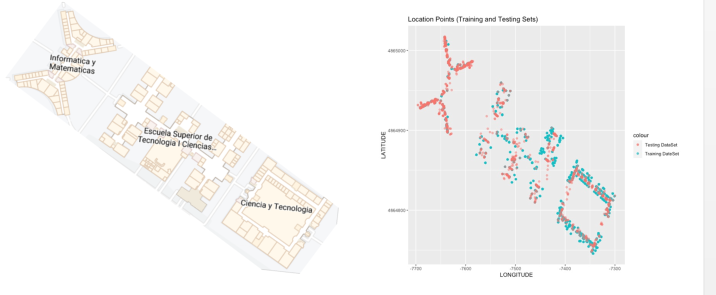
*Figure 3*

## Attribute Information

* **Attributes 001 (WAP-001) – 520 (WAP-520):** negative integer value from -104 (*extremely weak signal*) to 0 (*extremely strong signal*) and +100 (*not detected*).
* **Attribute 521 (Longitude)**: longitude. Negative real values from -7695.94 to -7299.79
* **Attribute 522 (Latitude)**: latitude. Positive real values from 4864745.74 to 4865017.36
* **Attribute 523 (Floor)**: altitude in floors inside the building. Integer values from 0 to 4
* **Attribute 524 (BuildingID)**: ID to identify the building. Measures were taken in three different buildings, categorical integer values from 0 to 2.
* **Attribute 525 (SpaceID)**: internal ID number to identify the Space Categorical integer value
* **Attribute 526 (RelativePosition)**: Relative position with respect to the Space (1-inside, 2-outside). Categorical integer value.
* **Attribute 527 (UserID)**: user identifier. Categorical integer value.
* **Attribute 528 (PhoneID)**: android device identifier. Categorical integer value.
* **Attribute 529 (TimeStamp)**: UNIX Time when the capture was taken.

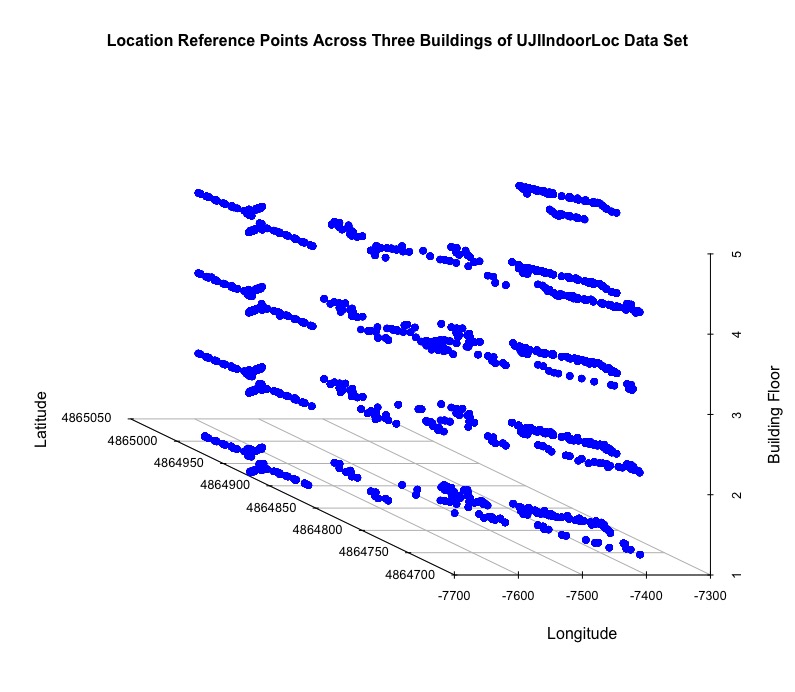
# Exploratory Data Analysis (EDA)

After the acquisition of the dataset and some preliminary pre-processing, the immediate actions were to further explore the data and dive deeper into the analysis in order to gain as much insight as possible before modeling. This approach and process not only helps to understand the information but identify patterns and other features that may not always be so transparent.

The initial approach investigated all three building as a whole/campus level. As shown in Figure 4, the plot created displays the location points for both the training dataset (red) and testing dataset (blue) of all three facilities. The WAP signal data that was captured (*right side of Figure* 4) shows a very consistent spatial pattern of the building layout as shown in the aerial view of the map (*left side of figure* 4).

*Figure 4*

To further investigate this information, a 3D-plot was created of each building in order to quickly identify the number of floors that were present in each facility. As shown in Figure 5, it is clear that building 2 contains 5 floors while building 0 and 1 only have 4 floors.



*Building 2*

*Building 1*

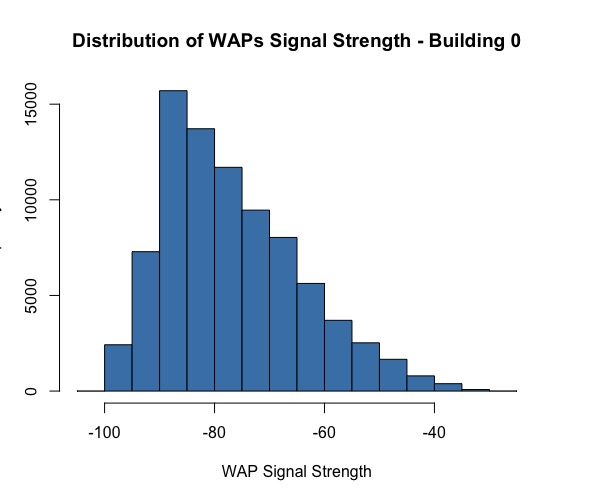
*Building 0*

*Figure 5*

Our goal was continue our EDA process and provide an additional layer of analysis at the individual building level. Our approached started at the higher level and then specifically targeted each building so we could capture as many details and insights as possible for each facility. The following sub-sections will highlight details for each building.

## EDA – Building 0

We explored the distribution of the WAP Signal Strength as seen in Figure 6. The majority of the signal detected are between -90 and -70 which is not relatively strong signals. Another aspect we wanted to investigate was the ‘Relative Position’ of the WAP signal capture for each building. Figure 7 takes an aerial view of the data of ‘Relative Position’ of inside the Space (red points) and outside the Space (blue points). An interesting observation is that most of the data signal detection are reflective of being ouside the Space (blue points) within Figure 7. To further investigate this observation, we took a closer look at ‘Relative Position’ by a floor by floor perspective, as seen in Figure 8.

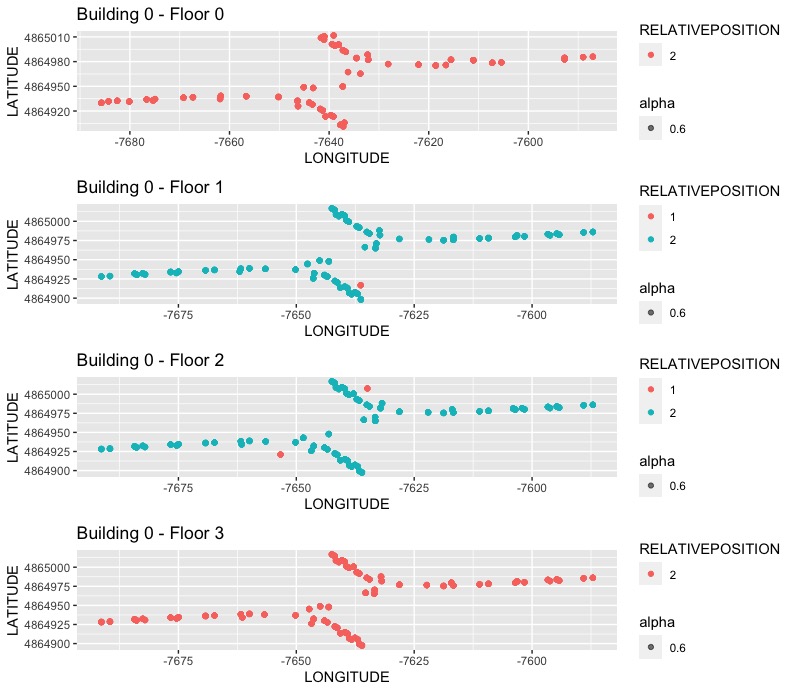
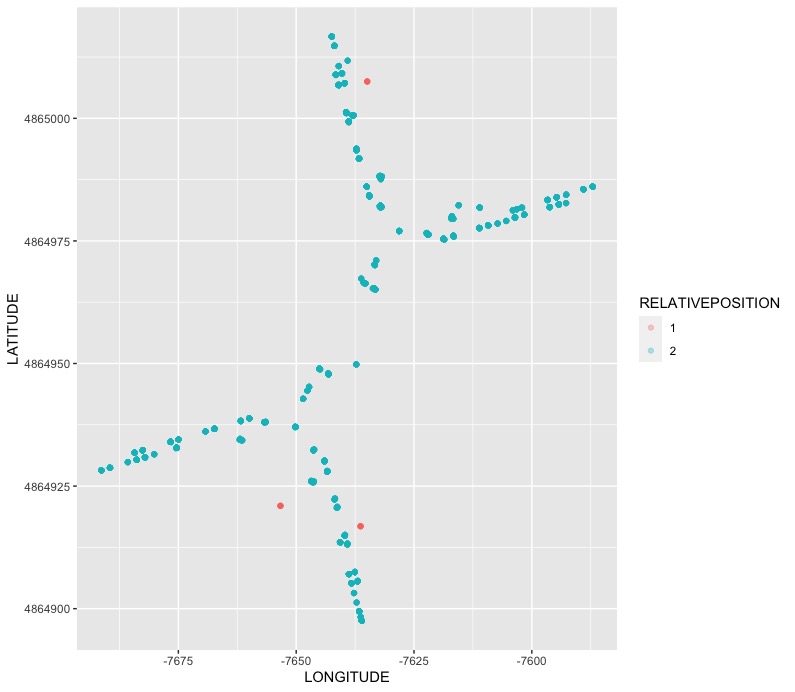
The following items were noted from Figure 8:

* ***Floor 0***: all observations are taken outside the Space (value of 2).
* ***Floor 1:*** *majority of observations taken outside the Space, very few signals inside the space.*
* ***Floor 2:*** *majority of observations taken outside the Space; very few signals inside the space.*
* ***Floor 3:*** *all observations take outside the Space*

The analysis conducted in Figure 8 confirms the initial observations in Figure 7 which most of the signal detections have a ‘Relative Position’ outside the Space, meaning in the hallways and not inside the actual room.

*Figure 6*

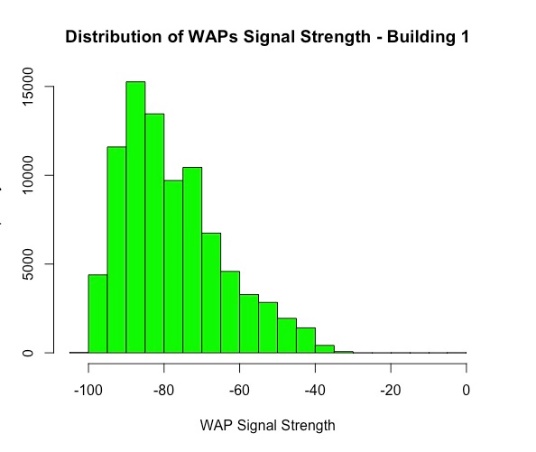
*Figure 7*



*Figure 8*

## EDA – Building 1

The distribution of signal strength was taken as seen in Figure 9. The distribution is similar to that of Building 0 with the majority of the data capture between -90 and -70s. To continue the ‘Relative Position’ analysis, Building 1 portrays as different mixture of inside and outside positioning in regards to Space when compared to Building 0, as shown in Figure 10. The initial observations from Figure 10 show heavier signal strength at different locations within the building, such as the right hand side versus the left hand side of the building.

To further explore this, a second plot was created that incorporates the floor data so that we could examine the individual floors of the building to confirm details. As seen in Figure 11, ‘Relative Position’ was portrayed by the various floors of Building 1. The following high-level observations are made from Figure 11. 

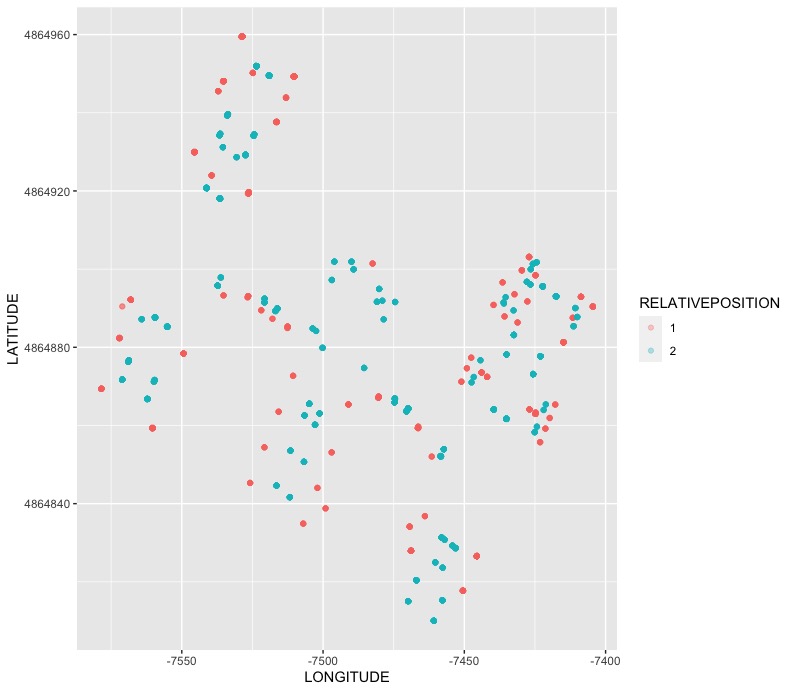
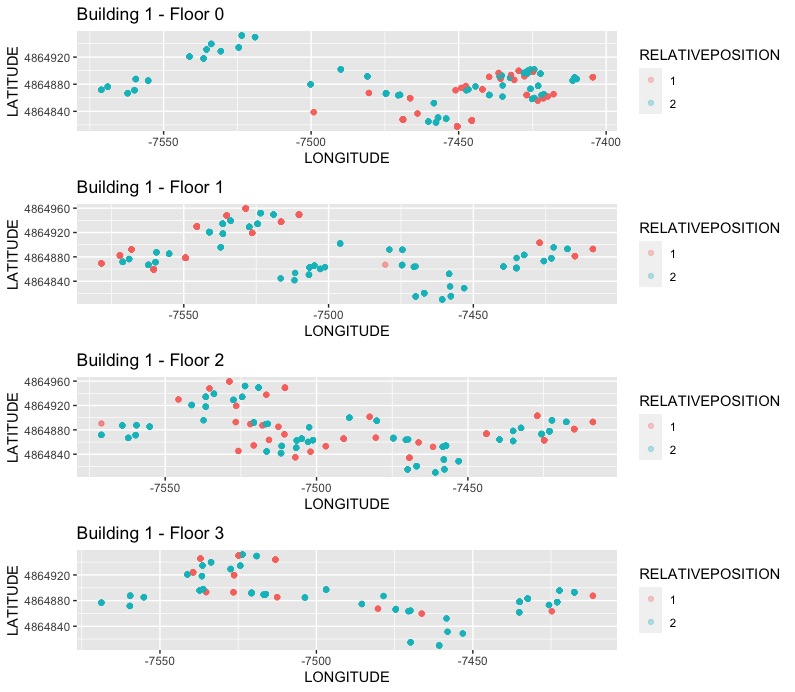
* ***Floor 0:*** *higher number of signal detection identified on the right hand side of the building.*
* **Floor 1**: *the left have side of the building captured more signal detention when compared to floor 0.*
* **Floor 2**: *contains the highest level of signal detection and has a balance mixture of inside and outside Relative Position across the entire floor*

*Figure 9*

* **Floor 3**: *signal detection is sporadic with more outside relative position detected*

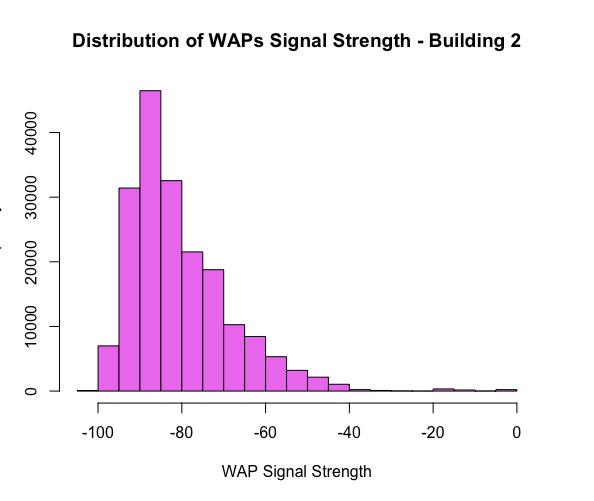
*Figure 10*

*Figure 11*



## EDA – Building 2

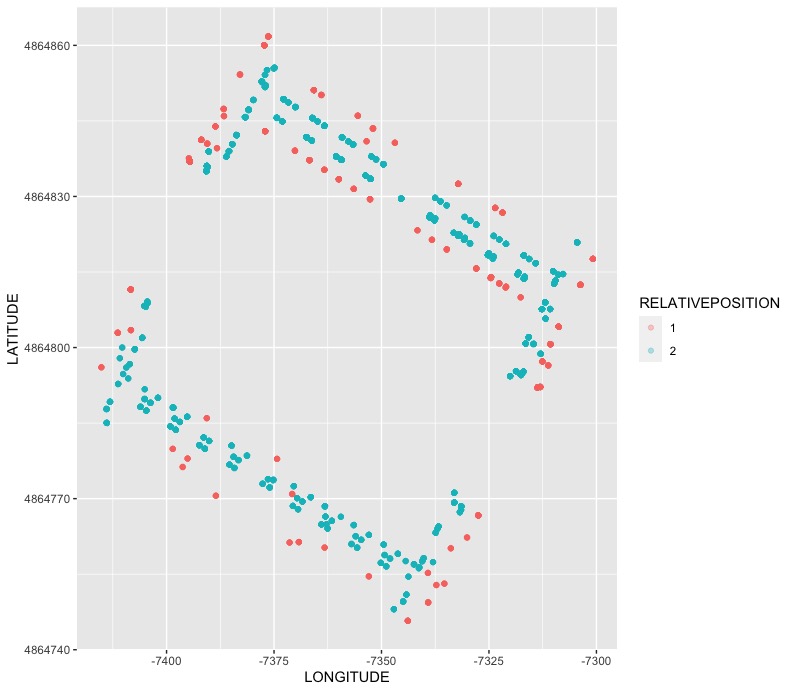
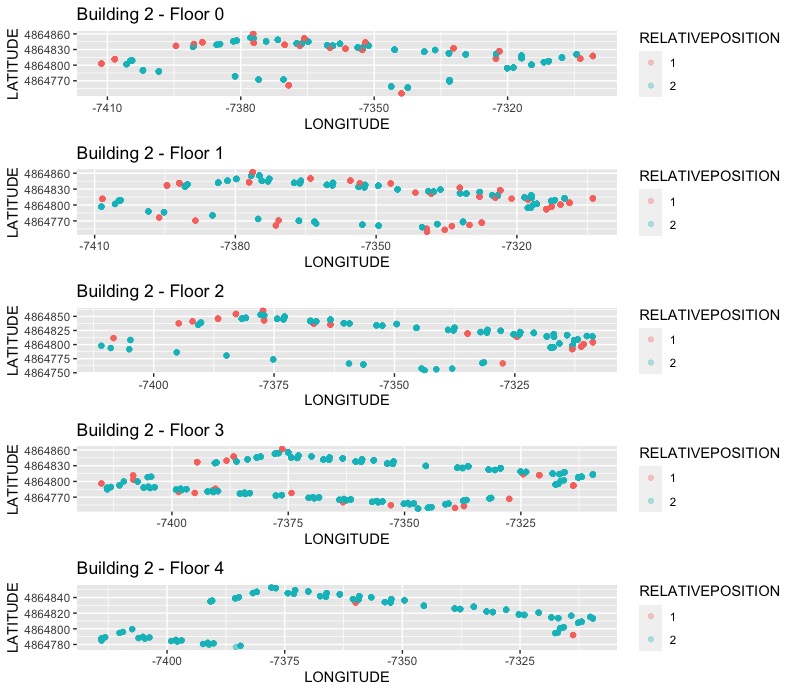
The distribution for Building 2 continues to follow a similar pattern distribution as Building 0 and 1, as seen in Figure 11. As with the prior to buildings, the same ‘Relative Positioning’ plot was examined. Based off Figure 12, the layout of signal detection is as expected with all indoor ‘Relative Position’ (red points) representing detection on the outskirts of the facility; while all outside detection (blue points) representing the corridors of the building. To keep a consistent method, we explored each floor of building 2. As seen in Figure 13, the signal detections are sporadic between floors.



* **Floor 0:** *one side of the building seen to detect*
* **Floor 1**: *a more consistent signal detection is observed with a mixture of indoor and outside relative position detection.*
* **Floor 2**: *like floor 0, limited signal detection is observed with the majority of the detection being outside the space.*

*Figure 11*

* **Floor 3**: *a more consistent spatial pattern is observed but more of the relation position is outside the space.*
* **Floor 4**: *sporadic detection is observed at various locations of the floor but most detection is outdoor the space.*



*Figure 13*

*Figure 12*

# Data Preparation

Coming out of the initial exploratory data analysis, a few items were identified for further processing. The following briefly describes the actions taken to prepare the data for modeling.

* **Removed all WAPs (columns) and records (rows) with zero variance detection**: if a WAP did not detect a signal or if a record had no detention signals recorded, those columns and rows were removed from consideration.
* **Feature Engineering – new combined feature**: a new LocationID field was created based off BuildingID, Floor, and SpaceID. This new feature is to serve as the ***dependent variable*** for modeling.
* **Removed features**: removed features from consideration that deem to have no value to the modeling; those fields include:
  + Longitude
  + Latitude
  + Floor
  + SpaceID
  + RelativePosition
  + UserID
  + PhoneID
  + TimeStamp

# Modeling

The goals were to evaluate three (3) different predictive models for the best performing outputs. The classification models taken into consideration included:

* KNN (k-nearest neighbor)
* Random Forest (RF)
* C5.0

The dependent variable used for these models was ‘LocationID’ that represented a combination of BuildingID, SpaceID, and Floor.

The approach we took was to build separate machine learning models for each separate building. The specifications is that when user were outside any building they would rely on the existing technology of GPS to locate their location. When the user transitioned from an outdoor environment to an indoor environment, the technology used would be the WiFi fingerprinting based of WAP signal detection. As mentioned, the approach was to create separate models for each building.

## Data Split

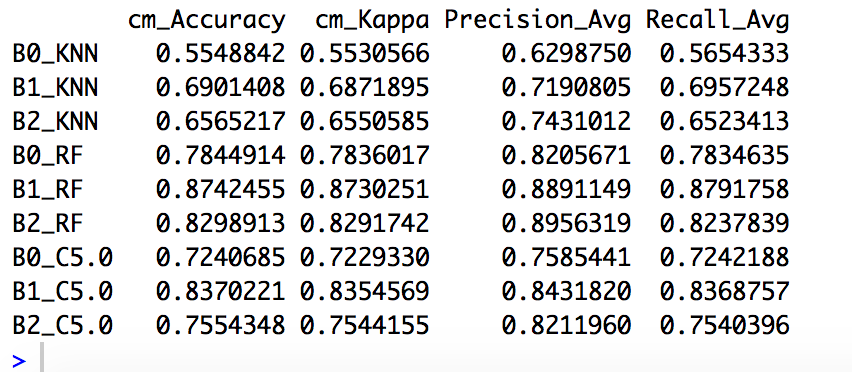
The data was split using and 80/20 rule where 80% of the data was used for training the model and the remaining 20% was used for testing that model.

## Performance Metrics

In order to evaluate the model performance, four key metrics were used for comparison.

* Accuracy
* Kappa
* Precision
* Recall

From the performance results below, the blue shaded box highlights that best performing metrics for each building. The buildings are denoted as B0-Building 0; B1-Building 1; and B2-Building 2. The ***Random Forest model is the best model*** that clearly out performs the other two models.



Accuracy and Kappa are performance measures used to compare and select the best fit model. Confusion matrix, post-resample and resample methods are also used to evaluate models.

* Kappa score is a metric that compares an Observed Accuracy with an Expected Accuracy and it is used not only to evaluate a single classifier, but also to evaluate multiple classifiers when they have been used on the same problem. In general it is less misleading than simply using accuracy as a metric; computations of Observed Accuracy and Expected Accuracy is integral to comprehension of the Kappa score, and is most easily seen in the use of a confusion matrix
* Observed Accuracy is simply the number of instance that were classified correctly throughout the entire confusion matrix.
* Expected Accuracy is defined as the accuracy that any random classifier would be expected to achieve based on the confusion matrix. The Expected Accuracy is directly related to the number of instances of each class combined with the number of instances that the machine learning classifier agreed with as being ground truth.

## Recommendation for Best Model

Based off the performance metrics, the Random Forest classification model is the best choice. In regards to all the four key metrics, Random Forest out performance the other two selected models for consideration. One item to note, it was observed that the cost performance (time to execute the model) for Random Forest seem to take longer, so that an area for consideration.