1. Task Description

Hi,

First of all, good job on analyzing the energy consumption dataset. The client was clearly impressed with your work.

Now it's time to begin a new project for a new client. Our client is developing a system to be deployed on large industrial campuses, in shopping malls, et cetera to help people to navigate a complex, unfamiliar interior space without getting lost. While GPS works fairly reliably outdoors, it generally doesn't work indoors, so a different technology is necessary. Our client would like us to investigate the feasibility of using "Wi-Fi fingerprinting" to determine a person's location in indoor spaces.

Wi-Fi fingerprinting uses the signals from multiple Wi-Fi hotspots within the building to determine location, analogously to how GPS uses satellite signals. We have been provided with a large database of Wi-Fi fingerprints for a multi-building industrial campus with a location (building, floor, and location ID) associated with each fingerprint. Your job is to evaluate multiple machine learning models to see which produces the best result, enabling us to make a recommendation to the client. If your recommended model is sufficiently accurate, it will be incorporated into a smartphone app for indoor positioning.

This is a deceptively difficult problem to solve. I'm looking forward to seeing what you come up with. After completing your analyses, please prepare an internal report on the project for IOT Analytics; because this initial report will be delivered to an internal audience, it can contain more technical detail than the report we will eventually present to the client. I have attached some sample data for you to use for your analysis.

Kathy

VP, IOT Analytics

2. Task Solution

Two data sets (trainingData.csv and validationData.csv) were provided by the CTO, to conduct a feasibility analysis to use Wi-Fi fingerprinting to determine a person's location in indoor spaces. The feasibility study will be conducted using three machine learning algorithms, which are: KNN, Random Forest and Xtreme Gradient Boosted Tree. The main goal is to predict the location (longitude, latitude and floor) of a user, based on WAPs signals.

2.1 Models Configuration

Model 1 – K- nearest neighbor (KNN)

This model contains the following processes and parameters:

- Pre-processing
- Data Partition (70/30)
- Train Control (method = "cv", number = 5, verboseIter = TRUE)
- Method = knn
- Predictions
- Post Resample

Model 2 – Random Forrest

This model contains the following processes and parameters:

- Pre-processing (normalization, dummy variables)
- Data Partition (80/20)
- Train Control (method = "cv", number = 5, verboseIter = TRUE)
- Method = rf
- Grid = expand. Grid (mtry=c(32))
- Predictions
- Post Resample

Model 3 – Xtreme Gradient Boosted Tree

This model contains the following processes and parameters:

- Pre-processing (normalization, dummy variables)
- Data Partition (80/20)
- Train Control (method = "cv", number = 5, verboseIter = TRUE)
- Method = xgbTree
- Predictions
- Post Resample

Algorithm Selection Criteria:

Table 1 contains the selection criteria for the algorithms used to asses the Wi-Fi fingerprinting feasibility.

Criteria	KNN	Random Forest	XGB Tree
Faster convergence	X		X
High accuracy	X	X	X
Used previously	X	X	X

Table 1 - Algorithms Selection Criteria

2.2 Data exploration

The figures below, are graphical representations of the Train and Validation data provided, in terms of position (e.g. Latitude and Longitude).

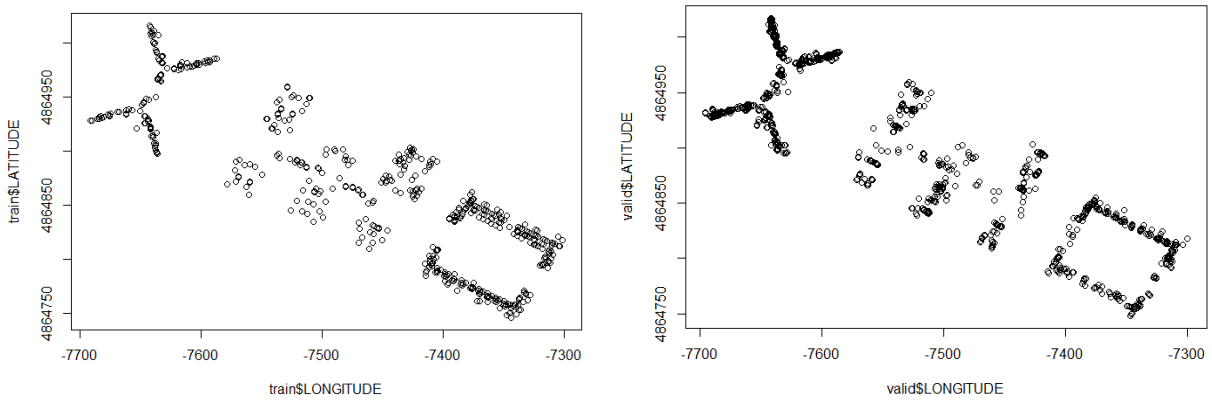


Figure 1 – Training and validation data sets

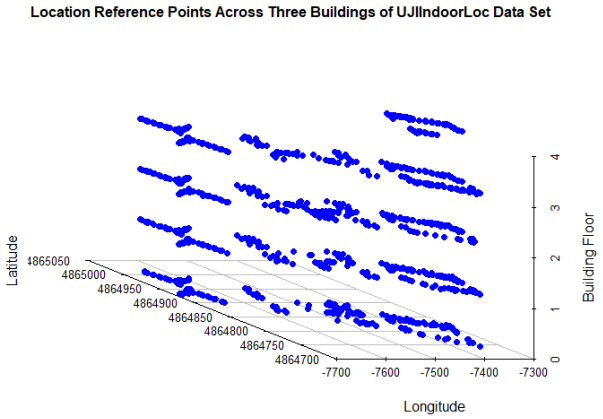


Figure 2 – Location Reference points across three buildings

In previous plots we can see that in validation set there are some areas that have no instances. In order to have a better picture of the data, both datasets will be merged and then divided in train and valid datasets.

2.3 Data pre processing

The flow chart listed below (Figure 3) is the data preprocessing steps defined to subset the data and evaluate the different features and algorithms.

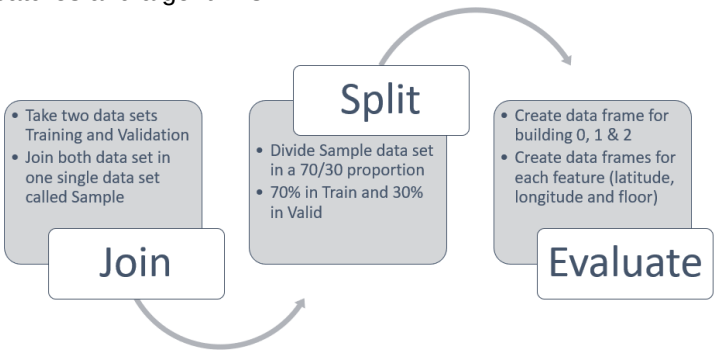


Figure 3 – Preprocessing steps

3. Results

The KNN, RF and XGBT models were fitted under the specified conditions, the model with the best performance is the Random Forest. Figures from 4 to 7, illustrates the fitting outcomes of latitude, longitude and floor accuracy features for three models.

Figure 4 shows the lowest RSME value for longitude (4.2) and latitude (3.3) in building 0, while the highest R2 value is 0.99 in latitude position for building 0; these values indicates a better fit.

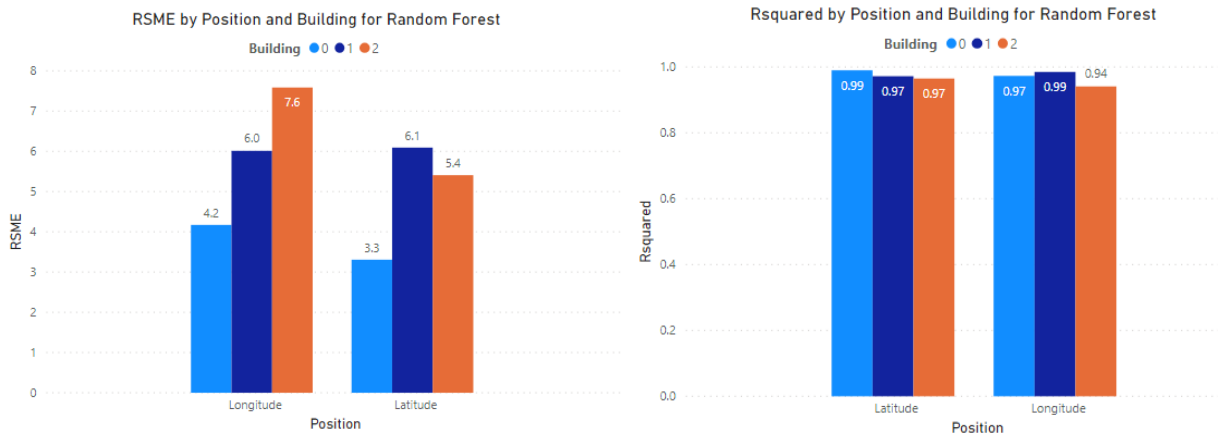


Figure 4 – RMSE and R^2 values for Random Forest model

Figure 5 shows higher RSME values when Random Forest and XGBT are compared, for example the lowest RSME for longitude is 5.9 and latitude 5.0 in building 0, while the highest R^2 value is 0.98 in latitude position for building 0. When these values are compared with RF results, the RF model has a better fit.

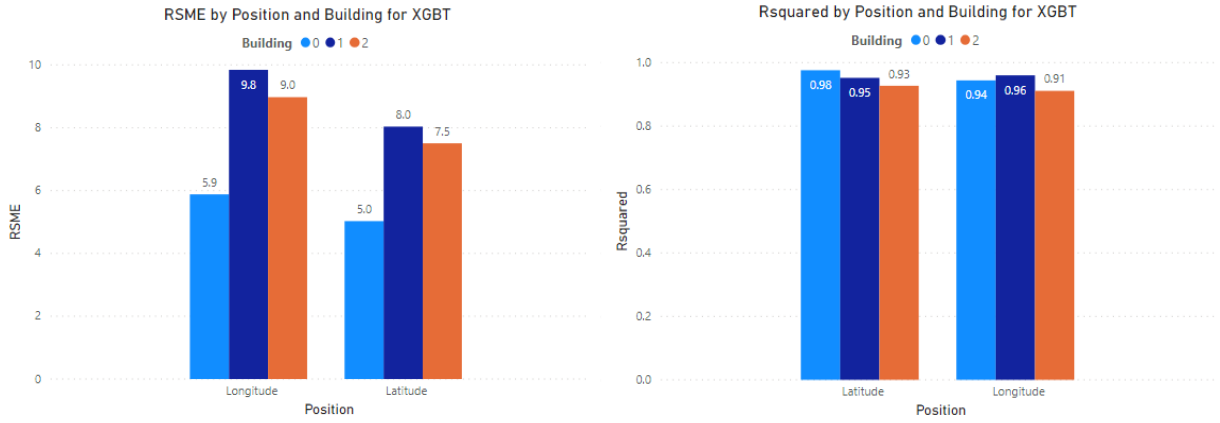


Figure 5 – RMSE and R^2 values for Xtreme Gradient Boosted Tree

Figure 6 shows higher RSME values when Random Forest and KNN are compared, for example the lowest RSME for longitude is 6.0 and latitude 4.6 in building 0, while the highest R^2 value is 0.98 in latitude position for building 0. When these values are compared with RF and XGBT results, the RF and XGBT models have better fit.

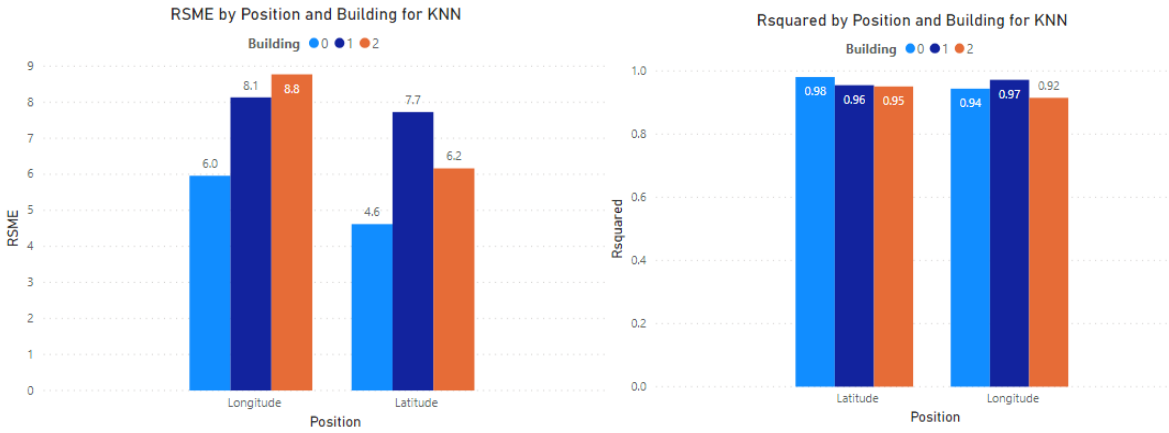
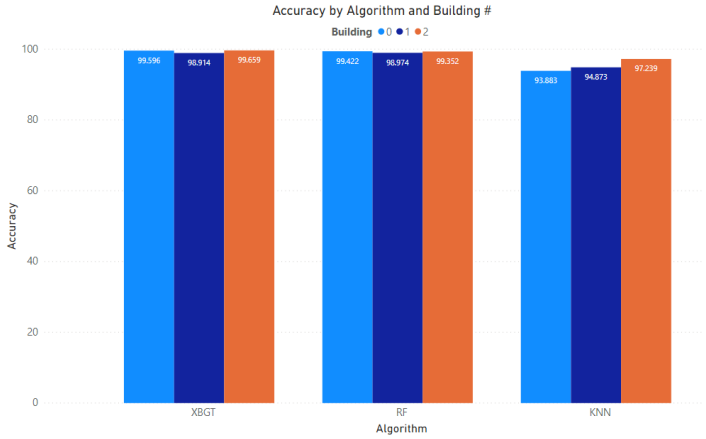


Figure 6 – RMSE and R^2 values for KNN model

Figure 7 shows the floor accuracy by building, in this image the XGBT model has better results against random forest and KNN. The highest accuracy is in building 2, while the lowest accuracy is in building 1.



		Building #		
		0	1	2
KNN	Accuracy	0.927	0.943	0.967
	Kappa	0.902	0.924	0.957
Random Forest	Accuracy	0.991	0.986	0.989
	Kappa	0.988	0.982	0.986
Xtreme GBT	Accuracy	0.995	0.989	0.996
	Kappa	0.993	0.986	0.995

Figure 7 – Floor accuracy for KNN, XGBT and RF algorithms.

Based on the previous outcomes, it's clear that random forest model shows the lowest RMSE and highest R^2 values in all buildings for latitude and longitude, while the best floor accuracy is for XGBT model.

However, the model selected to predict positioning for all the features, will be the random forest models. Figure 8 illustrates the sequence of steps defined to predict the positioning with random forest algorithm.

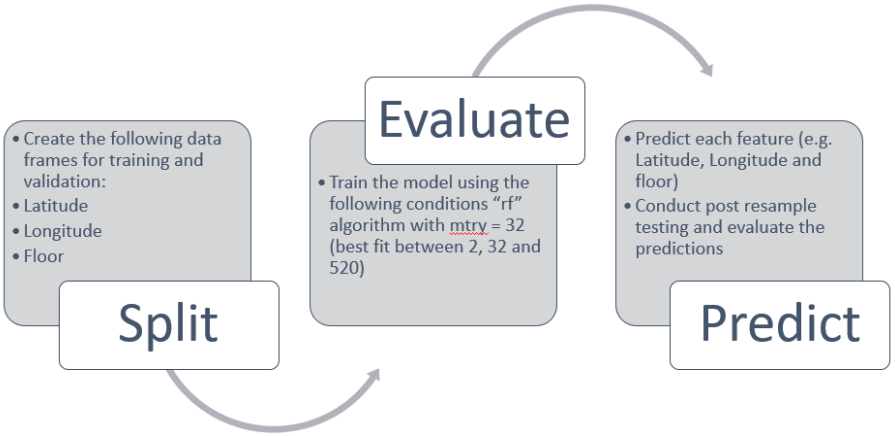
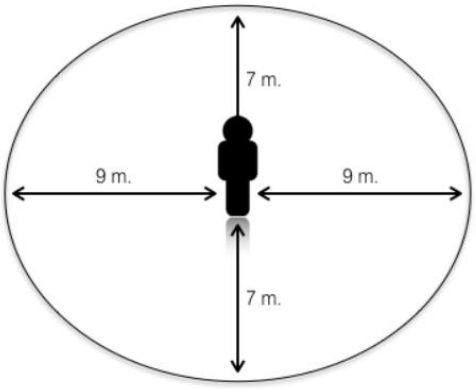


Figure 8 – Workflow of Final Model fit

The outcomes of the predictions for Latitude, Longitude and floor accuracy shows that predicted error has been around 7 meters to north and south, and 9 meters to west and east.



	Latitude	Longitud
RMSE	6.921	8.990
R-squared	0.989	0.994
MAE	4.078	4.899

Floor Accuracy	99.14%

Figure 9 – Random Forest Prediction

This error range is suitable for predict de position in indoor spaces, taking into account that the error range of GPS (location system in outdoor) goes from 3 to 15 meters, depends on the quality of the appliance.

The post resample testing for Latitude shows lower variance and high percentages of accuracy higher than 93% for RMSE, R-squared and MAE.

	Prediction	Post Resample	Variance	% of Accuracy
RMSE	6.921	6.45	0.466	93.27%
R-squared	0.989	0.991	-0.002	99.80%
MAE	4.078	3.824	0.254	93.77%

Figure 10 – Post Resample for Latitude

The post resample testing for Latitude shows lower variance and high percentages of accuracy higher than 94% for RMSE, R-squared and MAE.

	Prediction	Post Resample	Variance	% of Accuracy
RMSE	8.990	8.822	0.168	98.13%
R-squared	0.994	0.994	0.000	100.00%
MAE	4.899	4.652	0.247	94.96%

Figure 11 – Post Resample for Longitude

In reference to floor feature we can see which floor has been the best and the worse predicted:

Floor #	Prediction	Confusion Matrix	% of Accuracy
0	1401	1378	98.36%
1	1661	1648	99.22%
2	1378	1370	99.42%
3	1555	1545	99.36%
4	330	330	100.00%

Figure 12 – Floor Accuracy with Kappa: 0.986 and Accuracy: 0.989

The best predicted has been floor 4, and the worse predicted floor 0. Maybe is because in floor 0, we have more users outside or in front of the door, so the signal has not been taking correctly.

4. Results Discussion

The goal of this project was accomplished, the feasibility to determine / predict the location of a user based on WAPs signals, was successfully achieved.

The predictions of the models are very accurate and quite representative. However, we have to take into account that we combined both datasets (Training and Validation) to create the models, since the training data had a lack of representativity. Hence, it is difficult to gauge if our model can have a problem of overfitting.

This model it's an accurate and inexpensive way to determine an user location in indoor spaces, several use cases can be developed to take advantage of the model.

5. Recommendations

Use the model RF to conduct the predictions since the R-squared values were the highest one and RMSE was the lowest one.

6. R scripts

Provided in a zip file