Indoor positioning system using Wi-Fi Fingerprinting

CAPSTONE PROJECT

Selvameenakshi Viswanathan
GROUP C- PART TIME DATA ANALYTICS
MENTOR: MR. E.LOGAN CRAWFORD

Background

Precise indoor positioning systems and indoor location tracking is becoming more common in today's connected and digitized world. While GPS is ideal for locating devices outdoors, it cannot be used for indoor localization because of loss of GPS signal inside buildings. That is where alternative location technologies such as Wi-Fi positioning come in.

Wi-Fi positioning uses already existing infrastructure and Wi-Fi access points (WAPs) to calculate where a device is located. Wi-Fi fingerprinting uses Received Signal Strength Indicator (RSSI) from multiple Wi-Fi hotspots within the building to determine location, analogously to how GPS uses satellite signals.

Business Objective

The goal of this project is to evaluate multiple machine learning models to predict the location of a person inside a building using Wi-Fi fingerprinting. The best model will then be recommended to client and if sufficiently accurate, it will be incorporated into a smartphone app for indoor positioning.

Dataset description:

This project uses UJIIndoorLoc Data Set from UCI Machine learning repository. . The database covers three buildings of Universität Jaume Campus. The Source data has 21048 observations out of which 19937 records are used as reference and 1111 records are used for validation.

Dataset Attributes:

Columns	Description
WAP 001- WAP 520	Denotes the intensity of the signal from each of the 520 wireless access point for a location. Has values ranging from -104 to 0. 100 means no signal is detected
Latitude	Latitude of a location
Longitude	Longitude of a location
Building Id	Building number- Possible values 0,1,2
Floor	Floor number – Possible values 0,1,2,3
Space Id	ID to identify the space- Office, corridor, classroom etc. Possible values- 1 to 254
Relative position	Relative position with respect to space. Possible values 1,2. 1 - Inside and 2 -Outside in front of the door
User Id	User Identifier- Possible values 1 to 18
Phone Id	Phone Id- Possible values 1-24
Timestamp	UNIX time when the capture was taken

Data Evaluation:

1. The data type of all columns is checked
2. The dataset is checked for nulls and missing data. No nulls are found
3. No duplicates are found
4. Unique values of Building, floor, space id and relative position are checked to see if there are any invalid data.
5. Summary of all columns is checked to understand the mean, median, min, max values of each column

Feature Engineering

The dataset has 3 buildings- 0,1,2 . Building 0,1 has 4 floors and Building 2 has 5 floors. Each floor has many Space Id's associated with it. Space ID can take any value from 1-254. To predict exact location of the user, BUILDINGID, FLOOR and SPACEID are combined to form a unique identifier called Location_Id (dependent variable for our Classification problem). There are 731 unique combinations in Location_Id.

Example:

BUILDINGID	FLOOR	SPACEID	Location_Id
0	0	102	00102
1	3	1	131
1	2	16	1216
2	4	108	24108

The first 2 digits of the Location_Id denotes the building and floor. The remaining digits denote the Space Id.

Data Preprocessing:

1. For training purpose, only the records that are captured outside the room, i.e. RELATIVEPOSITION==2 is considered. This reduced the reference records from 19937 records to 16608 records.
2. The data type of dependent variable Location_Id is category.
3. After creating new Location identifier, BUILDING ID, FLOOR, SPACE ID and RELATIVE POSITION columns are removed
4. Irrelevant columns like LATITUDE, LONGITUDE, USERID, PHONEID and TIMESTAMP are also removed
5. Only WAP columns are selected as features.

Data Preparation for Modeling:

1. Modeling is done using 2 approaches. One is training entire dataset, and another is training individual buildings
2. 4 Samples are created. Sample 1 is entire data, Sample 2 is Building 0, Sample 3 is Building 1, Sample 4 is Building 2

Sample	Number of Classes in dependent Variable
Entire dataset	731
Building 0	256
Building 1	162
Building 2	313

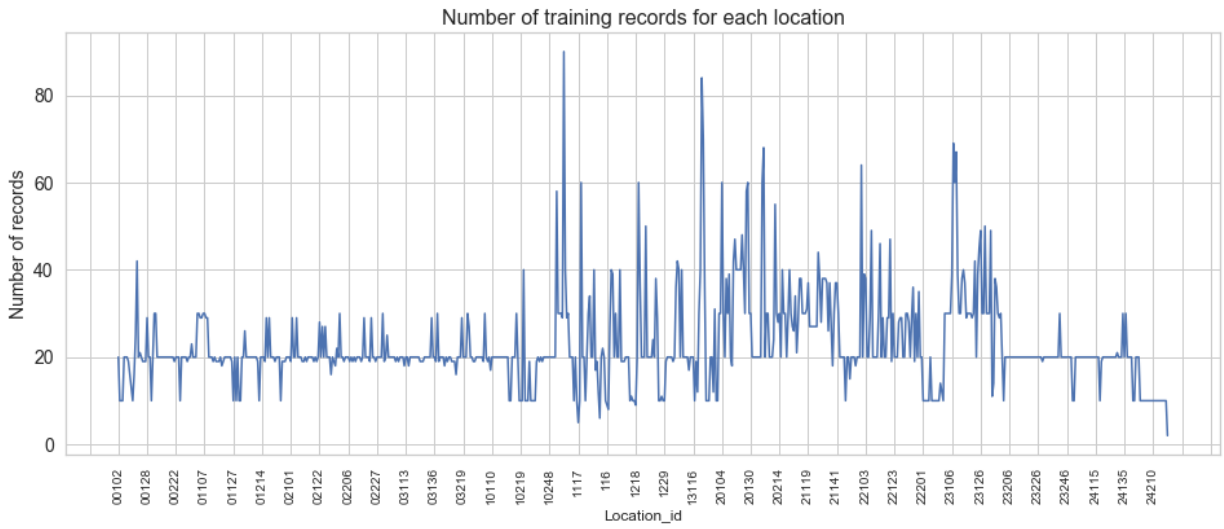
3. 2 datasets are created for each building sample.
    a. Out of the box dataset (OOB) without any feature selection. All 520 WAP's selected.
    b. Feature selection by eliminating zero variance columns (ZV)

Sample	Dataset	No. of rows	No. of features
Entire dataset	OOB	16608	520
Entire dataset	NZV	16608	445
Building 0	OOB	5220	520
Building 0	NZV	5220	200
Building 1	OOB	3559	520
Building 1	NZV	3559	187
Building 2	OOB	7829	520
Building 2	NZV	7829	188

4. Each of the dataset is then divided into training and test set (75:25 ratio) for modeling. Stratified train test split is used so that all classes are represented proportionally in test data.

Known issues with data

1. This is a highly imbalanced Classification problem with huge number of classes(731) and number of records per class ranges from as many as 90 to as less as 2 records per class. Below graph shows the number of records for each of those 731 locations in the source data.



2. Too many attributes may pose a challenging feature selection
3. Training the model may take time as it is a high dimensional data (520+ columns)

Modeling

The below algorithms are chosen to train the 11 datasets mentioned in the previous table.

1. Random forest (rf)
2. Decision Tree
3. K Nearest Neighbors (KNN)
4. Support Vector Classifier (SVC)

Comparison of Training Accuracy

For each sample, both the datasets (OOB,NZV) are trained by 4 algorithms. The NZV dataset performed better than OOB for most of the samples.

Comparison of OOB dataset of all samples:

Metrics	**Model**	**Entire dataset**	**Building 0**	**Building 1**	**Building 2**
Accuracy	**Random Forest**	0.664	0.676	0.7766	0.6081
	Decision Tree	0.434	0.45	0.5855	0.3765
	KNN	0.461	0.4342	0.5616	0.4272
	Support Vector(SVC)	0.563	0.5624	0.642	0.5245

Comparison of NZV Dataset of all Samples:

Metrics	Model	Entire dataset	Building 0	Building 1	Building 2
Accuracy	**Random Forest**	0.6688	0.68	0.7791	0.606
	Decision Tree	0.4399	0.4467	0.5909	0.3681
	KNN	0.4611	0.4342	0.5616	0.4272
	Support Vector(SVC)	0.5611	0.5614	0.6434	0.5253

Insight: Random Forest performed well for all the samples compared to other 3 algorithms in training

Comparison of Post resample metrics:

Random Forest is used to predict the test data as it was the top performing model in training. Classification report is used to get the performance measures like Precision, Recall, Specificity etc. for each class.

Since this is highly imbalanced classification problem, Precision and Recall are better metrics for comparing models. Precision and Recall are calculated for each class. Therefore, to compare models, Macro Precision and Macro Recall are calculated by taking average of all classes.

Comparison of OOB dataset Random Forest results:

Sample	Accuracy	Macro Precision	Macro Recall
Entire dataset	0.8205	0.86	0.81
Building 0	0.7678	0.79	0.77
Building 1	0.8921	0.89	0.87
Building 2	0.8049	0.89	0.80

Comparison of ZV dataset Random Forest results:

Sample	Accuracy	Macro Precision	Macro Recall
Entire dataset	0.8145	0.85	0.80
Building 0	0.78	0.80	0.78
Building 1	0.8842	0.88	0.86
Building 2	0.8227	0.9	0.82

Evaluate findings

Once Random forest is chosen as the best algorithm and post resample results of both OOB and ZV datasets are compared, decision must be taken whether to use single model for training all data or each building must be trained individually. To decide, we must compare the Recall results of each model using Classification report.

Recall for each class is the ratio of number of instances correctly classified to total number of instances in that class. We can then calculate how many classes are in each recall percentage bracket. For example, 0% recall means all those classes are not at all classified and 100% recall means those classes are always classified correctly.

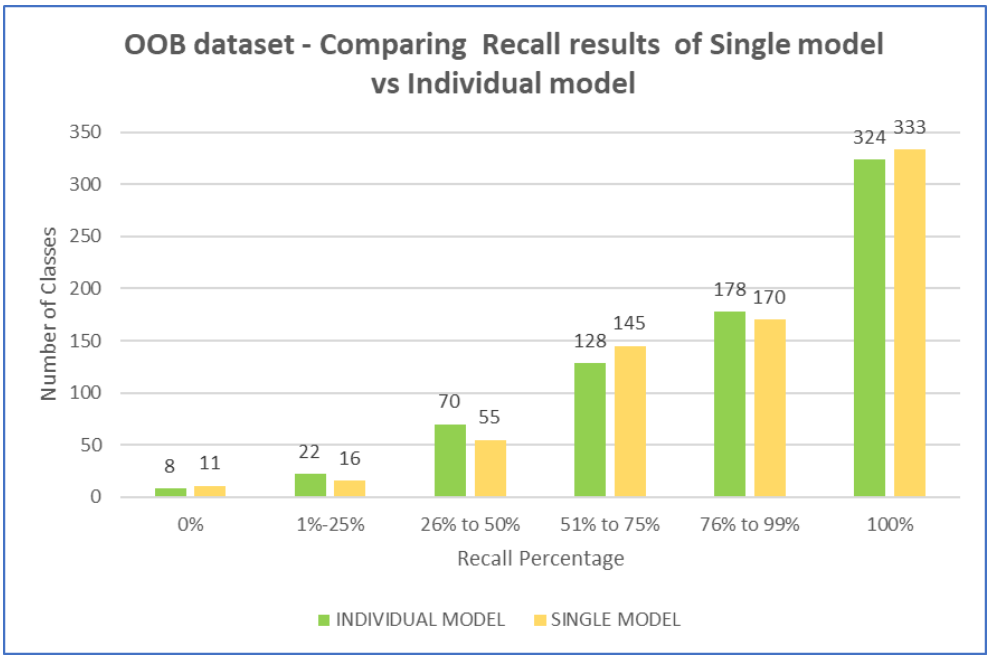
Below table shows the number of classes in each Recall percentage bracket for each sample when trained by Random Forest.

OOB dataset: Comparison of Random Forest results of Single model and Individual models

Recall Percentage	Building 0 Number of Classes	Building 1 Number of Classes	Building 2 Number of Classes	Total no. of Classes Individual model	From Single model
0	3	4	1	**8**	**11**
1-20	7	1	14	**22**	**16**
26-50	26	10	34	**70**	**55**
51-75	63	16	49	**128**	**145**
76-99	63	26	89	**178**	**170**
100	94	105	125	**324**	**333**
Total Classes	**256**	**162**	**312**	**730**	**730**

From the above table, we can understand that training ***OOB dataset*** by Single model is better than training by individual models as

1. The number of classes/locations that are 100% correctly classified are 333 by Single model and 324 by Individual models.
2. 648 (145+170+333) Classes out of total 730 classes have recall percentage greater than 50% while trained by Single model whereas only 630 classes(142+306) have recall percentage greater than 50% when trained by Individual models.



Single model performs better for OOB dataset.

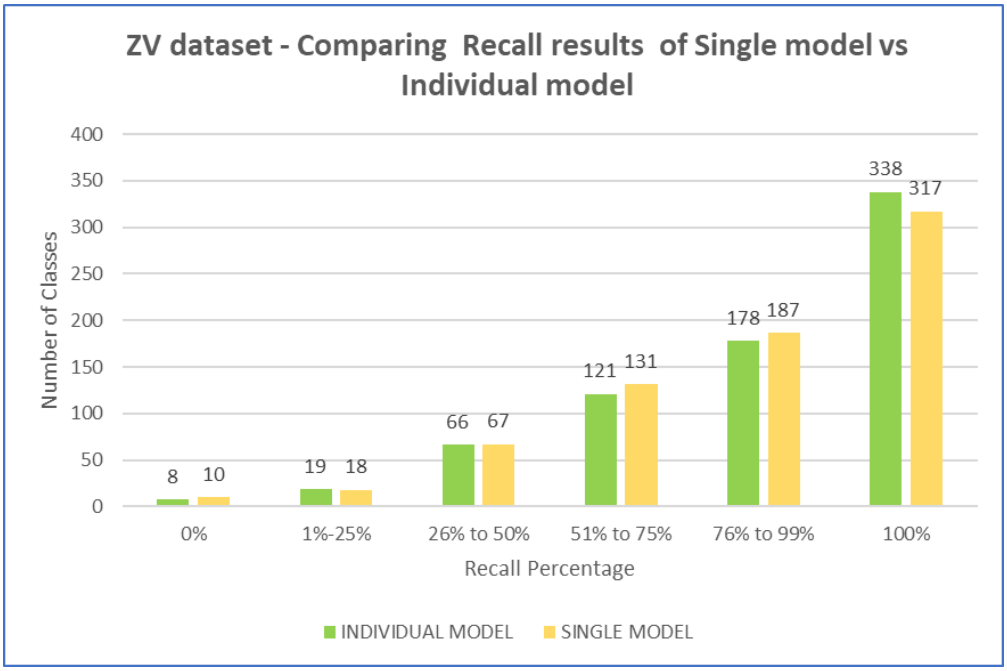
ZV dataset: Comparison of Random Forest results of Single model and Individual models

Recall Percentage	Building 0 Number of Classes	Building 1 Number of Classes	Building 2 Number of Classes	Total no. of Classes Individual model	From Single model
0	2	5	1	**8**	**10**
1-20	6	3	10	**19**	**18**
26-50	27	7	32	**66**	**67**
51-75	55	18	48	**121**	**131**
76-99	68	26	84	**178**	**187**
100	98	103	137	**338**	**317**
Total Classes	**256**	**162**	**312**	**730**	**730**

From the above table, we can understand that training models by individual building is better than training all buildings by a single model as

1. The number of classes/locations that are 100% classified are 338 for individual models and 317 for Single model

2. 516 (178+338) Classes out of total 730 classes have recall percentage greater than 75% while trained by individual buildings whereas only 504(142+306) have recall percentage greater than 75%



Individual model is selected for ZV dataset.

Model ,Dataset and Method Selection:

Random Forest is the chosen model. For choosing dataset and method, let us compare the recall results of OOB Single model and ZV Individual model results.

Recall Percentage	Single OOB model	ZV Individual model
0	11	8
1-20	16	19
26-50	55	66
51-75	145	121
76-99	**170**	**178**
100	**333**	**338**
Total Classes	**730**	**730**

Selected Method : Training Individual buildings

Selected Dataset : ZV dataset

Selected Model : Random Forest

Recommendations

After evaluating the findings, the recommendation is to ***train each building separately by Random Forest and use Zero variance dataset*** to get maximum number of locations correctly classified.

Sample	ZV dataset	Random Model	Test Accuracy	Recall
Building 0	zv_wifib0	zvRFfitb0	78.01%	78%
Building 1	zv_wifib1	zvRFfitb1	88.43%	86%
Building 2	zv_wifib2	zvRFfitb2	82.28%	82%

Recall Percentage range	Number of locations in that range
0	8
1-20	19
26-50	66
51-75	121
76-99	178
100	338
Total Classes	**730**

More than 70% of the locations(338+178 out of 730) are classified correctly with recall percentage greater than 75%.