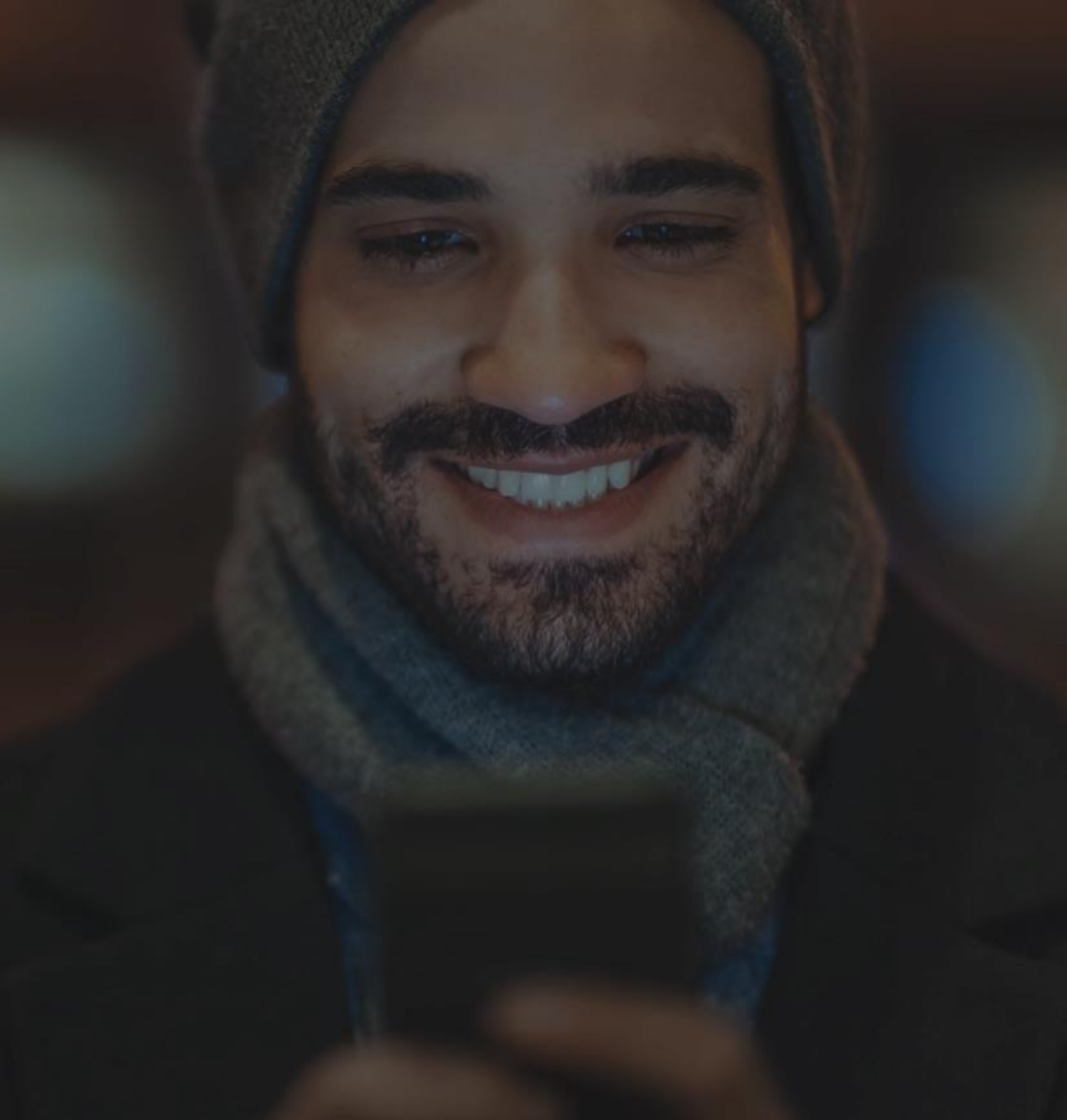
Indoor Locationing Predictions using Wifi Fingerprinting

IOT Analytics:
For Internal Use Only

Jennifer Brosnahan, MPH



Background

- Our client is developing a system to help people navigate a complex, unfamiliar interior space on a college campus.
- They would like us to investigate the feasibility of using wifi-fingerprinting to determine a person's indoor location.
- If a model meets or exceeds minimum specifications, it will be incorporated into a smart phone app for indoor locationing on a college campus.

Business Objectives (Goals)

1

Build multiple different models to predict a person's location in indoor college campus spaces using Wifi signals

2

Evaluate models to determine if one can be built that meets or exceeds client's minimum specifications (defined on next slide)

3

If a model is successful, deploy into smart phone indoor locationing app

Client Minimum Specifications

- Indoor location must be as precise as predicting within 10-15 feet of the indoor room, also defined as 'SpaceID' within source data. Relative position, or whether individual is outside or inside of room, is unnecessary.
- Performance metrics to determine if model is successful or not:
  - Accuracy and kappa scores on test data reaches at least 90% or higher
  - Precision (accuracy of minority class) on test data reaches at least 90% or higher
  - Recall (coverage of minority class) on test data is at least 90% or higher
  - F1 Score for multi-class problem achieves 90% or higher
  - Ideal performance metrics of model achieve results of 95% or higher
- Ideally, client prefers one algorithm to cover predictions on entire campus. If performance metrics cannot be achieved, consider multiple algorithm strategy.

Data Description

- Data was collected by 20 individuals using mobile phone devices on a college campus in Velencia, Spain
- The data source is the UJIIndoorLoc WLAN database
  - 19937 observations of 529 variables
  - 520 of the variables (98% of dataset) are homogenous wifi-access points providing numeric values representing signal strength
  - 9 remaining variables are Longitude, Latitude, Floor, BuildingID, SpaceID, Relative Position, UserID, PhoneID, and Time
  - Contains no missing values

Data Management Methodologies

Feature Engineering

- A unique dependent variable ('location') was engineered in order to pinpoint room location on campus by concatenating 3 features from out-of-box (OOB) dataset:
  - 'BuildingID' (3 buildings, values = 0, 1, 2)
  - 'Floor' (5 floors, values = 0, 1, 2, 3, 4)
  - 'Space ID' (125 total space IDs)
  - Results in 735 total unique 'location' classes

Feature Selection

- As this problem requires predicting location using Wifi Access Points (WAPs), all non-WAP variables were removed
- Removing zero variance variables utilized as a strategy to help reduce dimensionality from OOB dataset during model training

Data Management Methodologies

Sampling Techniques

- Because OOB dataset is extremely large, two sampling techniques were employed to create additional datasets from which to train models:
  - Sample 60% of floors (3 of 5 floors) – as devices have an ability to pick up wifi signals from adjacent floors, a dataset containing only floors 1, 3, and 5 was created
  - 40% random sample of OOB observations

Three primary datasets were developed from raw source data and utilized in model training process (outlined on following slide)

Dataset Descriptions

Dataset 1: Out-of-box (OOB)

- 19,937 observations and 521 variables
- 520 WAP variables retained as independent variables

Dataset 2: Floors 1, 3, 5 and feature selection

- Floors 2 and 4 removed from OOB dataset
- 9,887 observations and 416 variables
- 415 WAP variables retained after zero variance variables removed

Dataset 3: 40% Sample and feature selection

- 40% random sample of OOB dataset
- Zero variance variables removed
- 7975 observation of 447 variables
- 446 WAP variables

Algorithms

Three classification algorithms were chosen as follows:

- C5O – automatically prunes ineffective nodes and branches, useful in high dimensional datasets
- Random Forest – handles large and high dimensional datasets well and performs automatic dimensionality reduction
- KNN – trains much quicker than tree-based models and new data can be added easily

Post-resample Performance Comparison of All Models

C50 algorithm is top performer (with OOB dataset) across all post-resampling results

	C5O oob2	C5O Floors135	C5O 40% Sample	RF oob2	RF Floors135	RF 40% Sample	KNN oob2	KNN Floors135	KNN 40% Sample
Accuracy	**0.96659**	0.74916	0.64174	0.95168	0.81345	0.74261	0.73136	0.62143	0.49739
Kappa	**0.96652**	0.74825	0.64098	0.95158	0.81276	0.74205	0.73080	0.62001	0.49631
Precision	**0.99177**	0.79609	0.67820	0.98139	0.85353	0.80065	0.76493	0.70576	0.56050
Recall	**0.95532**	0.73306	0.59338	0.93900	0.79226	0.69135	0.70035	0.61217	0.44202
F1 Score	**0.97320**	0.76328	0.63296	0.95972	0.82176	0.74200	0.73122	0.65564	0.49426

Model Recommendation

- Recommend deploying C50 OOB algorithm for smart phone app
  - C50 Accuracy, Kappa, Precision, Recall, and F1 Score exceed 95% ideal goal
  - Model achieves high overall accuracy, but more importantly predicts minority classes excellently (Precision 99.2%)
  - Model covers minority classes very well (Recall 95.5%) compared to other algorithms
  - Of 3622 locations to predict in test set, model predicted 3500 correctly (3.34% error rate)
  - Model covers entire campus (vs. separate model for each building)

Result enhancement proposals

Although model exceeded minimum specifications, confusion matrix reveals most inaccuracies occurred in 5^{th} floor of building 3 with space IDs between 101 and 147. Building 2 floor 2 also showed some inaccuracies.

1. Build 3 separate algorithms each tied to a specific building to reduce number of classes. Doing so may help improve correct predictions within building 3, especially within space IDs 101 and 147
2. Install additional WAPs to help improve signal strength in weaker Wifi areas, such as in building 3, to add to source data, or replace weaker WAPs with stronger signals.
3. Try L1 or L2 algorithms, which are known for penalizing when wrong decisions are made in imbalanced datasets
4. Try an enhanced weighted K-nearest neighbor (EWKNN), which may help improve accuracy on imbalanced datasets by changing the number of considered neighbors

Alternative Solutions to Wi-Fi Fingerprinting

- Acuity Brands provides LED integrated indoor positioning light fixtures which send flickering patterns readable by phone receivers and link to indoor locationing maps. www.bytelight.com
- IndoorAtlas uses a variety of technologies for indoor locationing, including geomagnetic positioning, Wi-Fi signals, and Barometric pressure (vertical movement), all of which are read be a cell phone's sensors. www.indooratlas.com/positioning-technology
- iBeacon transmitters, or Bluetooth Low Energy (BLE) devices, project signals to nearby electronic devices which can be used to determine device's physical location. developer.apple.com/ibeacon
- Infrared (IR) systems use infrared light pulses to locate IR receiver signals installed in each room of a building. IR tag pulses emitted by receivers can be read by IR receiver device.
- Real-time fingerprinting apps rely on user collaboration to "check-in" and label fingerprints on their devices, which are then stored in a repository and pass through an algorithm that defines their current location. www.foursquare.com

Summary

- Out of several models built and datasets used to predict an individual's location based on wifi signals, the **C50 OOB model performed best**
- All performance metrics for C50 OOB algorithm exceeded minimum specifications and **met ideal goals established by client**
- Recommend **deploying C50 OOB algorithm** for Indoor Locationing smart phone app