EVALUATION OF TECHNIQUES FOR WIFI LOCATIONING

Analysis and Recommendations

Kevin Burr

January 2020

WHAT WE WILL COVER

- Goals of the Project
- Description and Location of Related Data Sources
- Data Management During Project
- Known Issues in Data and Proposed Resolution
- Algorithm Performance Summary and Recommendation
- Recommendations
- Summary Statements
- Lessons Learned

GOAL OF THE PROJECT

- Investigate the feasibility of using "wifi fingerprinting" to determine a person's location in indoor spaces.
- Evaluate multiple machine learning models to see which produces the best result.
- Provide a recommendation on best performing algorithm.
- Provide suggestion on how to improve the overall accuracy.
- Prepare an internal report on the project for IOT Analytics

DESCRIPTION AND LOCATION OF RELATED DATA SOURCES

- Provided a data set by client: UJIIndoorLoc.zip
- Contained two files of data: Training Data and Validation Data
- Training Data has 520 WAP variables, plus variables for Building, Floor, Space, Lat/Long and a couple more for more specific location once within or near a space
- Required for this project are the 540 WAP predictors and a single label that is a combined/concatenated value from Building + Floor + Space
- For the validation data there are 1111 observations with 529 variables
- For the training data there are 19937 observations with 529 variables

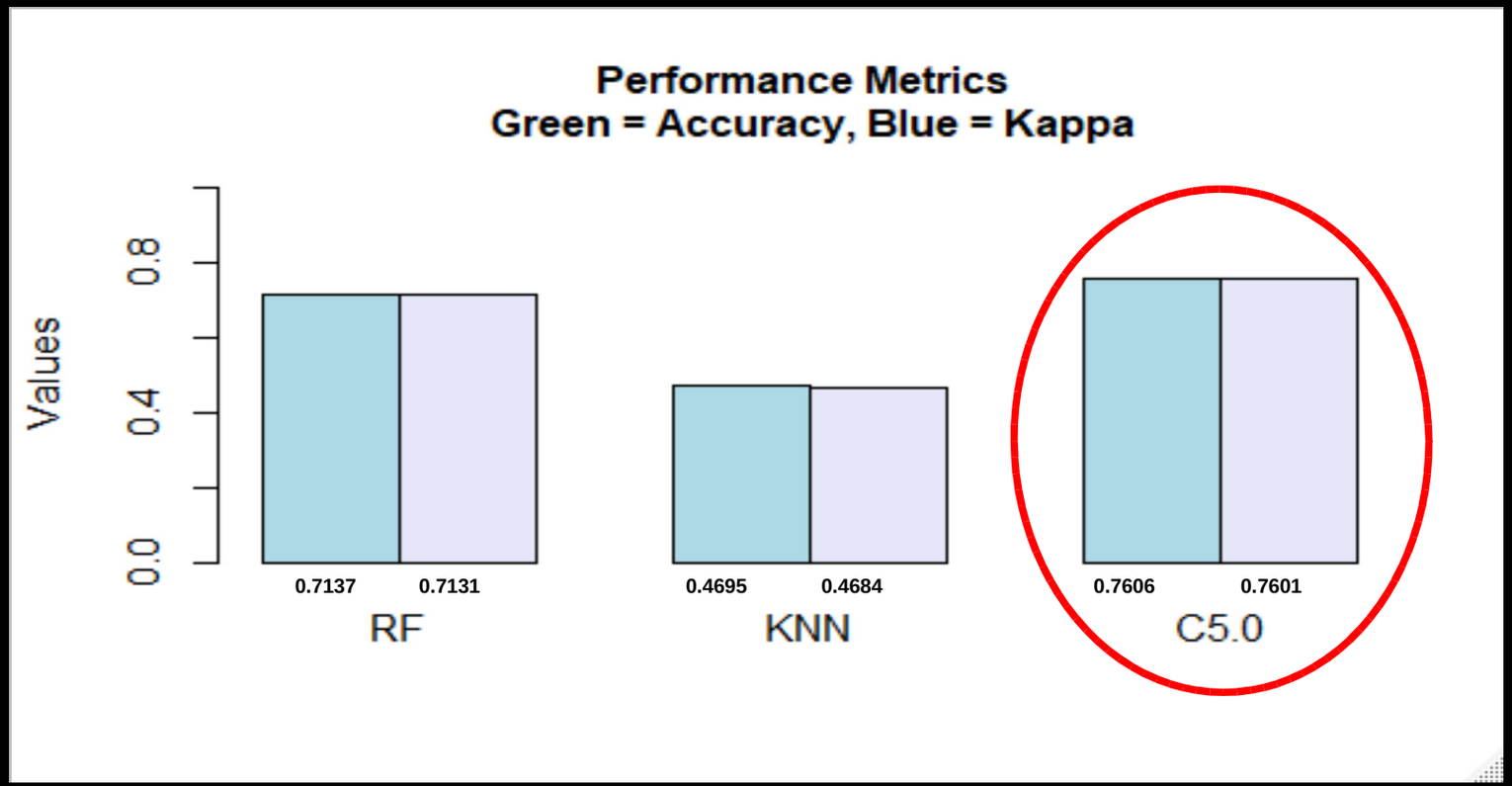
DATA MANAGEMENT DURING PROJECT

- Data will be managed on a single laptop during algorithm data wrangling and algorithm assessment work.
- Laptop is password protected.
- All data will be removed completely from the development laptop at conclusion of project.
- All results will be communicated via summary in various files and a single presentation. These artifacts will be communicated to stakeholders via company email or FTP as necessary.

KNOWN ISSUES IN DATA AND PROPOSED RESOLUTION

- The only real issue identified with the data is in the validationData file. There are NO spaceid values, which is one of the three primary location variables chosen to use in the training data to determine specific location of an individual based on cell phone relative signal strength index (RSSI).
- For the purposes of the approached to algorithm evaluation put forth in this effort, the following variables were not used in either the training or validation data sets: TIMESTAMP, USERID, PHONEID, RELATIVEPOSITION, LATITUDE, LONGITUDE
- Relative Signal Strength Index values read from the 540 WAP's (Wide Area Network Access Points) are between -100 to 0 indicating a positive reading and stronger the more negative, or 100 representing no reading at all. Since the values were all the same unit of measure, the data was not scaled or normalized prior to processing.

ALGORITHM PERFORMANCE SUMMARY AND RECOMMENDATION



Each Algorithm is a Classification Algorithm. Metrics used to measure performance for Classification Algorithms are Accuracy and Kappa.

C5.0 has the highest values for Accuracy AND Kappa of the three algorithms assessed.

C5.0 is the recommended algorithm for this project.

RECOMMENDATIONS

- Given the training phase of the algorithm development process included complete observations including values for the SPACEID variable, SPACEID values missing from the validation test reduce the overall accuracy of predictions for that data set. A new validation data set should be created with SPACEID values.
- Since in this case location information is based on WAP single strength, it may be possible to develop an app for mobile devices that send location information directly from a mobile device. This is an alternative method to the one devised for this assessment.
- It may be possible to derive more accurate location results with additional WAP's. An assessment of WAP coverage in the buildings could determine whether or not there are gaps in coverage which would result in weaker signals than truth would be otherwise – or if WAP signals overlap, there would be disruption between the two resulting in artificially weak signal strength.

RECOMMENDATIONS

- Additional processing power could allow for further evaluation of other models that took too long to run for this project timeline such as Gradient Boosted Trees and Support Vector Machine algorithms.
- It may also be possible to essentially feed results from one algorithm into another algorithm and derive a total overall higher level of predictive success (Ensemble Learning) – this may also result in lower success metrics as well. So, close evaluation would be interesting but needs to be assessed relative to the same accuracy and kappa metrics.
- It would be interesting to consider TIMESTAMP in further algorithm development. Since Time of Flight (ToF) measures an ack timestamp with the device at virtually instantaneous timing, with no real degradation of performance unlike strength of wifi signal that could be disrupted (weakened) by things like position of device near metal structures, or having to transmit through walls. A series of TIMESTAMP values paired with RSSI may provide more accuracy since TIMESTAMP enables creation of a timeline of movement of the device.

SUMMARY STATEMENT

The goal of this project has been met.

We arrived at a recommendation of a reasonably useful algorithm with accuracy commensurate with the missing key attribute – spaceid in the validation data.

We also arrived at a handful of interesting recommendations for business to consider in terms of different data or possible algorithm development/approaches that may yield improved results and should be at least considered in future rounds of algorithm development and evaluation.

All performance metrics and predicted results based on recommended algorithm were provided in separate files for stakeholders to independently evaluate recommended algorithm predicted results quality.

LESSONS LEARNED

- Given the size of the data set (15000 in training and 1100 in validation) with hundreds of predictors it is crucial to have a powerful compute system to process a reasonable amount of data in certain models in a reasonable amount of time. I tried to process 5000 records in XGB and SVM. Each algorithm ran over 7 hours and I had to terminate the process in order to make progress on this project. Its really important to ensure ONLY the minimally required amount of data that reasonably represents the data space is used to evaluate these models.
- Since some of my models ran for so long, I had to abort processing – in one case after over 7 hours of processing. It would be really interesting to know if those algorithms were more accurate than the three I chose due to processing time relative to amount of time for this project.
- It's important to ensure you review the quality of the data and understand the limitations of the data and how to either account for it in the data wrangling phase to create a valid training/test data set or how to account for the model quality impact of these data quality issues.
- It took me a lot of time to research the right models to run, how to structure the label, and how to create the charts and final summary data outputs. The good news there are TONS and TONS of really great online resources that answered all my questions. I just needed to have patience and spend the time doing the research.
- From a business knowledge domain perspective, I knew nothing about this problem and it was critical for me to understand my process and results to thoroughly research the business space first – and really throughout the time of running the project.

QUESTIONS

-