Methods for Analyzing a Wireless Network via Customer Mobile Application Data

By

Margaret K. Schweihs

A Capstone Project Paper Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Science

In

Data Science

University of Wisconsin –La Crosse

La Crosse, Wisconsin

TABLE OF CONTENTS

CAPSTONE SUMMARY

Course Outline/Description

Capstone is a course in which students develop and execute a project involving real-world data. Projects include: formulation of a question to be answered by the data; collection, cleaning and processing of data; choosing and applying a suitable model and/or analytic method to the problem; and communicating the results to a non-technical audience.

Course Objectives:

- Formulate a research question, problem or hypothesis that can be answered or tested using real-world data;
- Collect and manage data to devise solutions to their research question, problem or hypothesis;
- Select, apply and evaluate models, tools and methods to address their research question, problem or hypothesis;
- Interpret and assess their results and evaluate the limitations of their findings;
- Prepare a professional report of their work and effectively communicate their findings to a non-technical audience.

Executive Summary

Margaret Schweihs interned at a national telecom company during the summer semester of 2018 while a student in the Masters of Science in Data Science (MSDS) program at the University of Wisconsin, La Crosse. She currently has a degree in Mathematics (BS) and work experience in IT as an Operations Research Analyst and a Unix System Administrator. This project was done in conjunction with her internship role as an Analytical Engineer and serves as the Capstone experience of the MSDS program.

Background

Currently, telecom companies have limited data on the user experience of cell phone service. Most of their data comes from the cell towers and engineers know expected coverage quality experienced by users at various locations. Major mobile application vendors are using popular applications as a proxy for collecting user experience data and providing this to cellular companies, giving them a potentially very valuable insight into the user experience. One of the most interesting aspects of this data is the fact that it is geo-tagged.

Section 702 of the Telecommunications Act of 1996 prohibits telecom companies from collecting Customer Proprietary Network Information (CPNI) from subscribers without written consent (Telecommunications Act of 1996, 1996). The FCC includes customer location in their definition of CPNI (Federal Communications Commission).

Rationale

It will be highly relevant going forward to analyze customer data collected via mobile applications. Geo-tagged customer data is relevant and valuable to many industries. In the telecom industry, video streaming makes up the biggest segment of cellular traffic. The assumption is that it is a representative metric for the overall health and quality of service. It is also the metric that they are very invested in optimizing.

Purpose

This project was a data exploration project with the goal of determining possible use cases and methodology for analyzing aggregated customer data. The project benefits the telecom company by providing useful documentation and research into an emerging type of dataset. As use cases emerge, this documentation can serve to streamline visualizations and analysis by create a common framework for use within the organization.

Methods

The application vendor provided a frontend dashboard for visualizing and selecting data, which could then be downloaded in CSV format. The data went through a rigorous exploration process in order to be fully understood by the Analytical Engineers. Metrics from the data were compared to metrics from in-house data sources to gain an understanding of how well the application data represented the network. Analytical engineers cleaned the data and built a Shiny Application to visualize various metrics available in the data. They created a scoring methodology to assess the quality of service in a geographical area; they also used two-layer clustering techniques to highlight area of underperformance.

Results

The application created was a potentially insightful tool. It allowed a large dataset to be systematically visualized and areas of interest highlighted using clustering techniques across multiple metrics. This means that a Network Performance Engineer could quickly see areas of under performance without manually comparing multiple metrics across hundreds of towers in a geographical area. The methodology used in this project is widely applicable and could be proposed as a best practice for quickly filtering areas of interest when tuned to the industry specific metrics.

Since the project described here was the initial data exploration phase, the next phase was validation with Network Performance Engineers to determine if the predictions made by the app were insightful enough to continue building an enterprise level solution from the application vendors data. The next steps were included in the final deliverable to the telecom company.

Privacy and Confidentiality

Due to confidentiality agreements, the telecom company and application vendor are not referenced by name in the project documentation. Due to privacy concerns, employees of the telecom company are referenced only by their position title.

Approved Project Proposal

Project Description

Current marketing and business models can utilize geo-tagged customer data to optimize their services and products. I am interested in the application of several data science techniques to geo-tagged customer data. First, I will explore best practices for visualizing customer location based data. Second, I will explore the predictive power of geo-tagged data collected via mobile applications, with the goal to generalize which predictive algorithms, geo-statistical methods or best practices can be applied to geo-tagged customer data. My goal would be to explore various algorithms and determine which algorithms best leverage the geo-tagging. Finally, I will explore methodology in applying time-series analysis to geo-tagged data and determine if there is a need for special considerations.

Rationale

Currently, cellular companies have limited data on the user experience of cell phone service. Most of their data comes from the cell towers and engineers know what coverage users at various locations should experience. Google is using YouTube as a proxy for collecting user data and providing this to cellular companies, giving them a potentially very valuable insight into the user experience. One of the most interesting aspects of this data is the fact that it is geo-tagged.

It will be highly relevant going forward to analyze customer data collected via mobile applications. Geo-tagged customer data is relevant and valuable to many industries. Video streaming in general makes up the biggest chunk of cellular traffic. The assumption is that it is a representative metric for the overall health and quality of service. It is also the metric that they are very invested in optimizing.

Proposed Project Title

"Methods for Analyzing a Wireless Network via Customer Mobile Application Data"

Proposed Project Purpose

This is a client-based project for a major US telecom company to be completed in conjunction with a 10-week summer internship program. The company has its headquarters in Chicago, Illinois. This project benefits the telecom company by providing useful documentation and research into an emerging type of dataset. As use cases emerge, this documentation can serve to streamline visualizations and analysis by create a common framework for use within the organization.

Project Objectives

The following are objectives that I wish to accomplish by undertaking the proposed capstone project:

1. Provide best practices in creating a dynamic visualization that serves as not only a dashboard, but also an analytics tool utilizing geo-tagged data

2. Explore various algorithms to determine if there is predictive power in the current data. Through this process, determine which algorithms best leverage the geo-tagging.

3. Generalize geo-statistical methods or best practices that can be applied to geo-tagged customer data.

Interviewees

- Lead Analytics Engineer and former RF Engineer: He is also a project mentor assigned by the client.

- Principal Engineering Data Analyst: He is a highly technical analytics leader in the organization and subject matter expert in network-based time-series data.

Application of Data Science Concepts

For this project I expect to use concepts from DS740, Data Mining, and DS705, Statistics for Data Science, in drawing statistical conclusions and exploring/comparing predictive models. I will also use skills from DS710, Programming for Data Science, in programming a proof-of-concept dashboard for data visualization. When considering best practices with respect to customer data, I will include ethical considerations, leveraging concepts from DS760, Data Science Ethics.

Description of Final Document

The final document for this project will be a PowerPoint presentation to provide in-depth coverage of the objectives listed above, any findings and visualizations that the telecom company authorizes for release, and detailed examination of the topic.

TIMELINE DOCUMENT

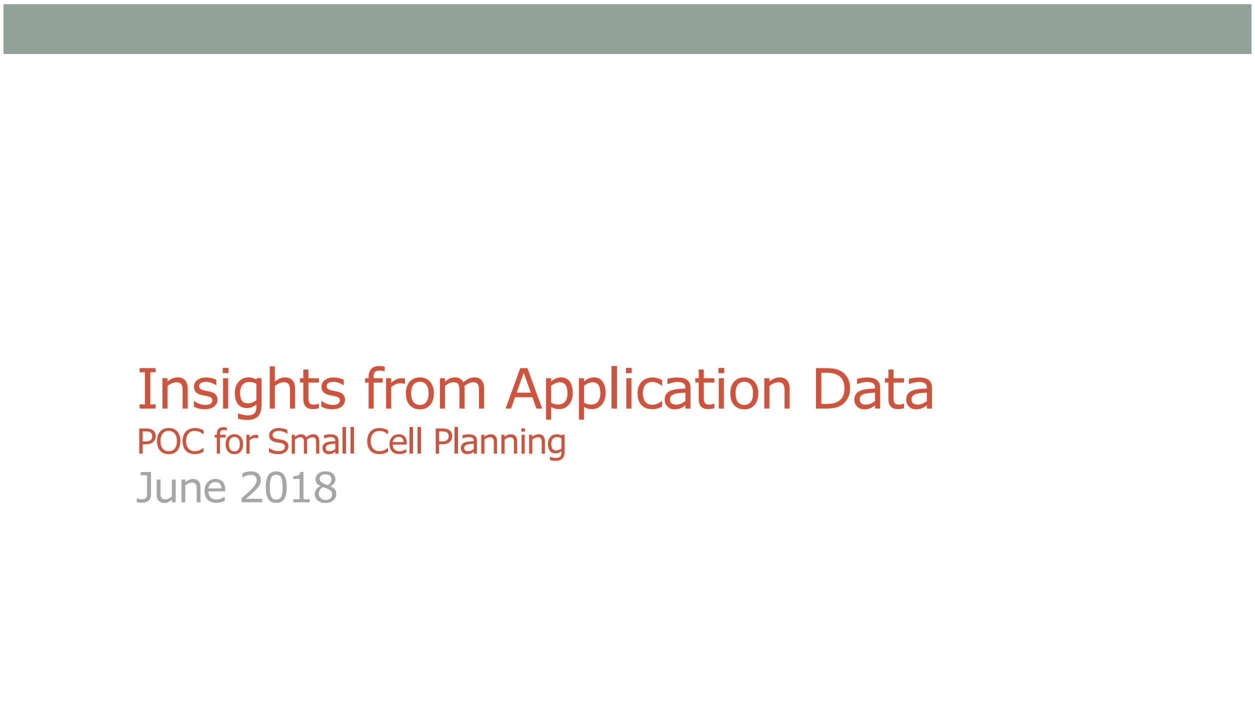
Week	Activities
Week 1	Receive data set and data dictionary. Begin initial data exploration and brainstorm use cases. Identify gaps in data and limits of the available data.
Week 2	Examine viability of data fields for focus on 1 use case: visualizing and predicting Video metrics based on radio frequency metrics. Report on the usability of the data for the use
Week 3	Focus on visualizing the data and determine which predictive algorithms are suited to the data. Begin documentation of methodology and standardize calculations.
Week 4	Focus on researching and applying best practices of data visualization for geo-tagged
Week 5	Focus on exploring predictive algorithms
Week 6	Generalize geo-statistical methods and best practices for dealing with customer data. Ethically, what should a data scientist pay attention to in order to ensure privacy is maintained?
Week 7	Explore methodology in applying time-series analysis to geo-tagged data.
Week 8	Work on documentation and refining deliverables
Week 9	Finish initial draft of project

Week 10 Refine Final Documentation for submission

Final Deliverable

The final deliverable is a PowerPoint presentation submitted as a separate file,

"InsightsFromAppData.pptx".



Insights from Application Data

POC for Small Cell Planning

June 2018

Project Reflection

The Capstone experience was very rewarding, due in no small part to the fact that I was interning with a great company. I have worked in corporate environments before and was able to hit the ground running with the team and the project. I gained exposure to enterprise level analytical work, used my prior work experience to work efficiently on the project, applied many concepts learned throughout the MSDS program, and even provided valuable training to my coworkers.

An advantage to working in an Enterprise level corporate environment was the access to lots of data science tools. The Analytical Engineering team had its own Red Hat Linux Server that hosted a Hadoop ecosystem, R Studio Server Workbench and connected to GitHub and the various data lakes in the company. This meant that you could directly query Hive tables from R Studio Server and that the R Studio Server had significant computing power. This was all very exciting to me, not only as a Data Science student, but also as a former Unix Systems Administrator.

In addition to a great computing environment, the company was using SCRUM methodology. I have never been in an “Agile” environment, but have heard about it as a useful project management system for software development. The team was using Jira software to manage the scrum workflows. Throughout the project, we entered stories and tasks in Jira to keep track of progress. We also organized our work into multiple two-week sprints. I enjoyed this workflow since it allowed the team to set realistic expectations, keep work delegated, and allowed freedom to manage your time autonomously at the same time. I think using this project management methodology contributed to our successful time management during the project.

As a former Systems Administrator, I was able to jump into the Hadoop environment immediately and start pulling and exploring data. I recognized several areas of improvement for the team and created an automated workflow for one of their reports by using scripting techniques in R. In doing so, I pointed out a few best practices that I hope will help them create more portable scripts in the future.

Throughout the internship, I applied concepts that I had learned in several of my MSDS classes. I used statistical analysis methods from DS 705, big data computing techniques from DS 730, data mining techniques from DS 740, and data ethics considerations regarding data privacy from DS 760. I appreciate the education that I have received throughout this program since I quickly realized that I had some very valuable skills and was able to make important contributions to the team.

One such contribution was regarding a training project that the team conducted shortly after the internship started. We were given several sets of data and a hypothetical scenario that an executive had read about a correlation between gas prices and cell phone data usage. In the scenario, we were to use the residuals to create a filter and extract the signal from the noise to determine the "true" correlation between the variables. This was an exercise exploring "spurious correlations". The Principal Data Analyst for the company designed the exercise. I helped my team create a well-documented R Markdown file that explained each step of the analysis and write two sample memos to the hypothetical executive explaining the true correlation and the recommended predictor variable, which was not gasoline price. This was a great training opportunity for my teammates and me.

In addition to the rewarding internship, the project itself was also very interesting. The two-layer clustering technique described in the final deliverable has the potential to be applied to many kinds of customer location data as a way of highlighting "hotspots" based on tuned parameters. One of the biggest challenges we faced with this data was the way in which it was aggregated. We worked with percentiles and averages rather than raw numbers. As I researched the best practices for releasing customer information, it was clear that aggregated data is the best practice and thus, there needs to be methodology for analyzing it. I think this concept could be explored further.

CAPSTONE DOCUMENTATION

ORGANIZATION MISSION/HISTORY

The client organization is a United States mobile network operator (MNO). “[The Telecom Company] is a regional carrier which owns and operated [one of the largest] wireless telecommunications networks in the United States, serving 5 million customers in 426 markets in 23 U.S. States as of the first quarter of 2017.” (Wikimedia, 2018) The company offers telecom services including individual and family service plans; business solutions; Internet of things devices and connection services for home and business; and business automation solutions (United States Cellular Corporation). In its 2017 annual report, the telecom company reaffirmed that its top business priority remains attracting new customers and protecting the current customer base; increasing revenues and profitability; and reducing costs throughout the organization (United States Cellular Corporation, 2017).

The telecom company uses data in three aspects of operations: marketing, customer care, and engineering. For marketing, the data is used for micro-segmentation and recommendation engines for cross sales and up sales; for customer care, big data provides a way to manage the customer experience; and for engineering, big data is being investigated for anomaly detection, automatic root-cause analysis, and expert systems for solution recommendations. (Principal Data Analyst, 2018)

According to an interview in January 2018 with the executive vice president and CTO of the telecom company, Kelly Hill reported that “[the company] is putting in place a comprehensive big data architecture and strategy for its business, with a deployment that has already begun and is focused on correlating network data with customer sentiment in order to predict and improve the customer experience, and shape how the operator approaches its infrastructure work,” (Hill, 2018). What began as a customer experience product has evolved into a massive data lake where all of the customer and network information can be ingested and potentially inform not only the customer experience, but also network optimization opportunities, according to the interview.

Service providers in each country have different rules and restrictions as to what kind of data can be exchanged through their network. There are numerous privacy and ethical considerations surrounding the use of telco data, some of which came to light in recent news headlines. On June 19, 2018, Sprint, Verizon, T-Mobile and AT&T announced that they would no longer sell real-time customer data. This arose from an incident where the third-party buyer, Securus Technologies, made a deal to release the real-time location information of cell phones to law enforcement without a court order. This deal violates the law and the Federal Communications Commission subsequently opened an investigation on the wireless carriers, prompting them to release the public statement. (Metz, 2018)

The law in question above is the Telecommunications Act of 1996, which prevents telecom companies from collecting location information about customers without their consent. This limits certain analytical studies that a telecom company can perform on its network. Thus, obtaining the information via a mobile application is the best way to gain this insight into the network.

People generally have to opt-in to “Location Services” and then read the privacy agreement before signing up for sites and apps that use location data. Individual data is rarely available in real time even for service providers. When operators provide subscriber information to third party companies, the current best practice is to anonymize the data by rolling it up into aggregations. For example, mobile phone network data can be aggregated at the cell tower level by considering the number of calls, Erlang, the number of SMS, the number of handovers, the number of location updates, and so forth (Calabrese, Ferrari, & Blondel). The Application Vendor in this project aggregated data by geographical bin, cell tower, and date range.

Client Communication/Correspondence

Most of the communication for this project was done via Skype teleconference, in-person meetings and Skype instant messenger. Additionally, internal emails from the telecom company are considered confidential and were not available for publication. Since I was an intern with the company, I was physically present in the office with team members, product owner and stakeholders throughout the project, therefore much of the relevant communication is undocumented.

Activity Updates

Activity Report # 1

Name: Margaret Schweihs

Project Title: "Methods and Best Practices for Analyzing a Wireless Network via Customer Mobile Application Data"

Agenda/goals: (Description of the work identified for this reporting period)

☑ Receive data set and data dictionary. Begin initial data exploration and brainstorm use cases. Identify gaps in data and limits of the available data.

☑ Examine viability of data fields for focus on 1 use case: visualizing and predicting Video metrics based on radio frequency metrics. Report on the usability of the data for the use case.

☑ Focus on visualizing the data and determine which predictive algorithms are suited to the data. Begin documentation of methodology and standardize calculations.

Resources and Investigation Methods: (software applications, library resources, journal articles, interviews, and other resources)

At this stage in the project, the main software tool has been R Studio Server. We have also utilized SQL developer to extract data from a database and we are currently formulating Hive queries to pull further data to supplement our analysis of application vendor's data. The application data is pulled via a web-based dashboard. In addition, we are using network analysis software called TrueCall to visualize the network as the carrier sees it today and compare it to the data that Google is providing.

I have read several ESRI white papers to gain an understanding of the "best practices" of geo-data. I have downloaded many articles relating to mobile data on a wireless network. My main focus right now is researching and applying predictive algorithms to the data.

Progress: (Project progress relative to the entire project):

I am 1/3 of the way through the initial plan that I proposed in the project proposal.

Achievements: (What was achieved based on agenda/goals identified above)

- ☑ Gained access to the application vendor dashboard to pull data
- ☑ Initial analysis revealed that we may have a solution for one of our use cases. We created a system of clustering data points that show "poor" service. When looking at these clusters in relation to the current network, it is possible that the clusters indicating potential future "small cell" sites OR opportunities for optimization.
- ☑ Thoroughly documented initial methodology and identified some crucial limitations of the data set.
- ☑ Created an interactive Shiny App Dashboard that visualizes the data, the predicted clusters, potential future cell sites, and current cell sites.

Questions: (Questions that came up during the reporting period that are directed to the instructor or reminders of questions that need further research)

- ☑ What machine learning methods are best suited for geo-data?

☑ How can I best present this information without divulging confidential network information?

☑ Still need best practices for customer geo-data

☑ Are the application vendor’s privacy thresholds too strict?

Next step: (Intentions for the upcoming reporting period what do you plan to do next?)

☑ Focus on researching and applying best practices of data visualization for geo-tagged data

☑ Focus on exploring predictive algorithms

☑ Generalize geo-statistical methods and best practices for dealing with customer data. Ethically, what should a data scientist pay attention to in order to ensure privacy is maintained?

Activity Report # 2

Name: Margaret Schweihs

Project Title: “Methods and Best Practices for Analyzing a Wireless Network via Customer Mobile Application Data”

Agenda/goals: (Description of the work identified for this reporting period)

☑ Refine presentation

☑ Clearly state next steps and possible directions of the project

☑ Conduct peer reviews with senior analysts and presentation reviews with product owner.

☑ Research the ethical and security considerations regarding customer geo-data

Resources and Investigation Methods: (software applications, library resources, journal articles, interviews, and other resources)

We are still using R studio Server for the main analysis tasks. We have used Hive queries to extract data from the data lakes. We also used TrueCall and the application data dashboard GUIs to visualize network information for our presentation.

I have interviewed two analytical professionals at US Cellular regarding data science and the project.

Researched boosting random forest models, logistic regression methods and association rules. I am currently researching two-layer cluster analysis for geo-data. A few of the sources I used are the following:

- Elements of Statistical Learning, by Jerome H. Friedman, Robert Tibshirani, and Trevor Hastie
- Spatial and Temporal Sentiment Analysis of Twitter data, by Zhiwen Song and Jianhong Xia
- Using TensorFlow Deep Neural Network to Classify Mainland China Visitor Behaviors in Hong Kong from Check-in Data, Shasan Han, et al.
- US Cellular AI Strategy, by Mario Vela

- ○ Documentation for arules and gbm packages in R

Progress: (Project progress relative to the entire project):

I am 2/3 of the way through the initial plan that I proposed in the project proposal. I think the bulk of the analytical work is complete and the project needs to be synthesized into a Capstone worthy report.

Achievements: (What was achieved based on agenda/goals identified above)

- ☑ Compare application data to in-house data to determine if we think we see a fair representation of our network in the provided dataset. We used in house network monitoring tools to pull raw data and compare distributions of variables.

- ☑ Examine and evaluate the assumptions that need to be made when dealing with aggregated data.

- ☑ Begin testing predictive data mining techniques to determine if we can predict areas of poor video resolution on network.

- ☑ Conduct peer reviews with senior analysts and presentation reviews with product owner.

- ☑ Used a two layer clustering method that would be really interesting to research further as a potential "best practice" in dealing with customer geo-data

- ☑ For DS780, wrote a 5-page paper on Mobile Network Operator data as it is collected from customers and used by marketers. This paper dives into the ethical and privacy concerns regarding customer geo-data and is relevant to this project as well.

Questions: (Questions that came up during the reporting period that are directed to the instructor or reminders of questions that need further research)

☑ What machine learning methods are best suited for geo-data?

☑ How can I best present this information without divulging confidential network information?

☑ Still need best practices for customer geo-data

Next step: (Intentions for the upcoming reporting period what do you plan to do next?)

☑ Present our initial Proof of Concept to the Stakeholders

☑ Determine "Next Steps" from project stakeholders

☑ Create a plan for moving forward with next steps and train someone on the project to take over after my internship period ends.

☑ Begin turning the project into a Capstone report. (Determine what information products from the telecom company that I can use)

Activity Report # 3

Name: Margaret Schweihs

Project Title: "Methods and Best Practices for Analyzing a Wireless Network via Customer Mobile Application Data"

Agenda/goals: (Description of the work identified for this reporting period)

☑ Presented our initial Proof of Concept to the Stakeholders

☑ Determine “Next Steps” from project stakeholders

☑ Begin turning the project into a Capstone report. (Determine what information products from the telecom company that I can use)

☑ Research the ethical and security considerations regarding customer geo-data

Resources and Investigation Methods: (software applications, library resources, journal articles, interviews, and other resources)

We are still using R studio Server for the main analysis tasks; PowerPoint for presentation and R Shiny for demo of dashboard.

Progress: (Project progress relative to the entire project):

I am still about 2/3 of the way through the initial plan that I proposed in the project proposal. The bulk of the analytical work is complete. I am working on gaining permission to use various graphs, charts and PowerPoint slides that were created during out exploration and analysis of the data. I am also trying to see if there is a similar open source data set available in order to write a little bit more specifically about the machine learning methods that I experimented with. If I cannot find a dataset, I still have a lot of room to explore methods and best practices without detailing any specific findings.

Achievements: (What was achieved based on agenda/goals identified above)

- Refined the presentation to convey the challenges posed by the aggregated data. Data aggregation is a best practice when anonymizing data, but it causes some "washing out" of metrics. For example, instead of getting specific RSRQ measurements, we got percentiles of measurements as observed in an aggregated bin. So, instead of one value, you have 5 values representing the percentiles. This proved to be a tough analytical challenge because you cannot infer the distribution of that variable

- Presented slide deck and demo of dashboard to two key stakeholders. They were excited by the analytical work that went into the presentation; however the feedback was that it was not ready for the next step: presentation to director and senior VP level, and needed further verification by the RF engineers familiar with our area of interest.

- For DS780, wrote a 5-page paper on the company's Customer Experience Management software implementation that strives to ingest customer data and network data and tune machine learning algorithms to predict anomalies and outages before they are reported.

Questions: (Questions that came up during the reporting period that are directed to the instructor or reminders of questions that need further research)

- How can I best present this information without divulging confidential network information?
- I have a request in through my manager to the legal team at the telecom company to get clearance for use of sterilized data products. I am looking into alternative data sources to

use for further examinations of the customer geo data concepts, but I think I have a good amount of analysis done if such a data set is not available.

Next step: (Intentions for the upcoming reporting period what do you plan to do next?)

☑ Begin solidifying research and writing the final report

CONTACT LIST

- **Analytical Engineer**, contacted via skype, email and phone:
    - He worked on every aspect of project with me, providing a sounding board for ideas and working to refine our dashboard and presentation as we receive feedback.
- **Project Mentor, Analytical Engineer,** contacted in person:
    - The project mentor is an Analytical Engineer on the team. He is a former Radio Frequency Engineer and has helped guide the direction of our analysis by explaining what would be most useful for network planning. He has served as a "Data translator" due to his subject matter expertise on the network metrics.
- **Principal Data Analyst,** contacted in person and via skype :
    - The Principal Data Analyst for the company has the high-level business perspective, leads the company's AI/ML initiative, and is very well versed in analytical methods. During this reporting period, he and the Lead Analytics Engineer provided a peer review of the work. They were able to point out things in our slide deck that would distract a non-technical audience from the "data story". He also had a few suggestions to improve our methodology.
- **Lead Analytics Engineer**, contacted via skype and email:
    - He provided feedback during a peer review of the project
- **Senior Manager of Analytics**, contacted via skype, email and in person:
    - He is the team manager and the product owner. He met with the other Analytical Engineer and I several times over the course of the project. We presented several iterations of our product (dashboard) and analysis and used his feedback to understand the direction the project should take.

- **Director or Analytics and Automation**: skype conference, final presentation 10 July 2018
  - He is the director and my manager's manager. The other Analytical Engineer working on the project and I presented our slide deck and proof of concept to him in order to gain feedback and determine the direction of the next steps of the project.
- **Lead Strategist, Tools & Data Systems**: conference, final presentation 10 July 2018
  - The Tools Architect finds and analyzes tools to determine if they should be implemented in the company's toolbox. He found the Application data/ dashboard that we are analyzing and was at the final presentation in order to gain feedback.
- **Application Vendor**, contacted via Google Hangouts led by the tools architect:
  - We had a videoconference with a project manager to better understand some of the limitations we were seeing in the dataset. We also gained insight into how other wireless carriers were using the data.

BIBLIOGRAPHY

Calabrese, F., Ferrari, L., & Blondel, V. D. Urban Sensing Using Mobile Phone Network Data: A Survey of Research. *ACM Computing Surveys , 47* (25:1), 25:20.

Federal Communications Commission. (n.d.). *Customer Privacy*. Retrieved 2018, from FCC: https://www.fcc.gov/general/customer-privacy

Hill, K. (2018, January 10). CTO and Executive Vice President. (K. Hill, Ed.)

Kaye, K. (2015, October 26). *THE $24 BILLION DATA BUSINESS THAT TELCOS DON'T WANT TO TALK ABOUT.* Retrieved June 22, 2018, from AdAge: http://adage.com/article/datadriven-marketing/24-billion-data-business-telcos-discuss/301058/

Metz, R. (2018, June 19). Verizon, AT&T, Sprint, and T-Mobile will stop selling your location information to data brokers. *MIT Technology Review , 2018* (June), p. 19.

(2018, June 29). Principal Data Analyst. (M. Schweihs, Interviewer)

Telecommunications Act of 1996. (1996). *110 Stat. 56* . Pub. LA. No. 104-104.

U.S. Cellular. (2018, 05 13). Retrieved 05 28, 2018, from Wikipedia: https://en.wikipedia.org/wiki/U.S._Cellular

United States Cellular Corporation. (2017). *2017 Annual Report.* Retrieved July 2018, from U.S. Cellular: http://s1.q4cdn.com/183458318/files/doc_financials/annual/2017/2017-Annual-Report-PDF-USM.PDF

United States Cellular Corporation. (n.d.). *Home Page*. Retrieved 07 11, 2018, from U.S. Cellular: https://www.uscellular.com

Wikimedia. (2018, 05 13). *U.S. Cellular*. Retrieved 05 28, 2018, from Wikipedia: https://en.wikipedia.org/wiki/U.S._Cellular

MISCELLANEOUS ITEMS

ANALYTICS METHODOLOGY AND CALCULATIONS

Data Sources

We utilize two types of data sets for this analysis. The first is the application data platform at either the cell sector or coordinate aggregation level. The second is data queried from the onsite containing cell tower location info.

Limitations of data sets:

- **RF Metrics** appear to be mostly complete across provided datasets at the coordinate level aggregation, having between 8 and 11% missing values across evaluated datasets. There also seems to be a decent amount of dispersion across values in these fields. At the cell sector aggregation level, these fields are intact with 0% NA values in LTE records.
  - RF Metrics tested:
    - LTE Signal Power Mean, LTE Strong Signal %, LTE Signal Power tiles, LTE RSRQ %tiles, RSRP quality RSRQ quality fields
- **Video Resolution Metrics** in the evaluated coordinate datasets contain a large amount of missing data ~95-99%. The cell sector aggregation dataset contains only ~8.5-12% missing data
- **Coordinate Aggregation Data** observations cannot be connected directly to a cell tower, given there are enough observations on each tower (lots of missing data).
- **Privacy Thresholds** currently prevent us from seeing areas with a small number of observations. As we increase the resolution of the data points, we lose information due to the privacy thresholds. Points that remain tend to be urban or points along roads.

Since we know that we lose Video Resolution Metrics at high resolution, we suspect that we are seeing a high occurrence of navigation data at high resolution.

Data Preparation and Standardization

Column Naming Conventions

Columns in the application data are renamed to reduce issues caused by special characters. All special characters including spaces are replaced with an underscore, "_". The following is the R code used to accomplish this:

```
names(mni_140m_df) <- gsub(" ", "_", names(mni_140m_df), fixed = TRUE)

names(mni_140m_df) <- gsub(".", "_", names(mni_140m_df), fixed = TRUE)

names(mni_140m_df) <- gsub(",", "_", names(mni_140m_df), fixed = TRUE)

names(mni_140m_df) <- gsub(")", "", names(mni_140m_df), fixed = TRUE)

names(mni_140m_df) <- gsub("(", "", names(mni_140m_df), fixed = TRUE)

names(mni_140m_df) <- gsub("%", "percent", names(mni_140m_df), fixed = TRUE)

names(mni_140m_df) <- gsub("__", "_", names(mni_140m_df), fixed = TRUE)
```

Location Filtering on Madison, WI

Our use case analysis focuses on the market in and around Madison, Wisconsin with the goal of capturing urban and rural data points. We bound the area of interest by latitudes in the range (-89.5, -88.6) and longitudes in the range (42.5, 43.2).

Removing rows with missing important values

We removed rows in which there was no latitude, longitude, LTE signal power mean, or RSRP Good RSRQ poor percent.

Calculations

Traffic Density

We were interested in taking traffic density into account when analyzing the application data. “Share of Observations” or “Observations %” is displayed in both the Cell Sector aggregation and Coordinate aggregation. This field displays the percentage of observations that occurred in the specified location with the specified user & network filters. The denominator is the total number of observations seen nationwide for a given operator across all user & network types. Since we were interested in the traffic density with respect to our area of interest, we used the following calculation to standardize this field after reducing the dataset to the area of interest:

Standardized Observation % = Original Observation % / sum of all Observation %s

R code:

```
mni_140m_s$observations <- mni_140m_s$Observations_percent /
sum(mni_140m_s$Observations_percent)
```

Each observation in the dataset makes up a very small % of the observations, even when limiting the data to the Madison area, we have ~30K observations at the 140m coordinate aggregation level we have ~1-2K observations. We scaled the variable in order to display it in our map visualizations:

```
rsize <- scale(mni_140m_s $observations)`
```

RSRQ performance metrics by cell

Color-coding RSRQ performance metrics in the map visualizations is currently done by quantile. For example, on the map of the 95th %-ile RSRQ (in dB), the values for the colors are determined by the 4 quantiles for the variable “LTE RSRQ 95th Percentile” and are colored from red to green, with red as the lowest quantile and green as the highest quantile.

		RSRQ Value		
		Poor	**OK**	**Good**
RSRP Value	**Poor**	Poor/Poor%	Poor/OK%	Poor/Good%
	OK	OK/Poor%	OK/OK%	OK/Good%
	Good	Good/Poor%	Good/OK%	Good/Good%

Weighted RSRP/RSRQ score

Application data displays the RSRP/RSRQ metric as an intersectional metric as follows:

For RSRQ:

- Poor = < -13 dB
- OK = -12 dB to -8 dB
- Good = > -7 dB

For RSRP:

- Poor = < -110 dBm
- OK = -109 dB to -90 dBm
- Good = > -89 dB

		RSRQ Value		
		Poor	**OK**	**Good**
RSRP Value	**Poor**	1	4	7
	OK	2	5	8
	Good	3	6	9

The data is displayed in 9 separate columns in the application dataset. The percent of observations that match each criterion is reported. In order to more efficiently compare and analyze RSRP/RSRQ values, we created a weighted RSRP/RSRQ score. We began by weighting each column from 1-9.

Then, we took the sum of the product of the value of each cell with the weight and stored this as the “Weighted RSRP RSRQ Score”. For example, for the row where the RSRP/RSRQ %-ile values are the following:

Poor/ Poor	**Poor/OK**	**Poor /**	**OK/Poor**	**OK/OK**	**OK/ Good**	**Good/ Poor**	**Good/ OK**	**Good/ Good**
21.1	17.8	0	13.3	45.6	2.22	0	0	0

		RSRQ Value (weights)		
		Poor	**OK**	**Good**
RSRP Value (weights)	**Poor**	21.1*1	17.8*4	0*7
	OK	13.3*2	45.6*5	2.22*8
	Good	0*3	0*6	0*9

We multiply each value by the weights and sum all 9 values to obtain the Combined RSRP RSRQ Score of 364.66.

This scoring methodology was identified as an area for future refinement. There may be a way to use Generalized Least Squares Regression to determine the weights that correlate the RSRP/RSRQ values to video resolution performance.

Clustering of LTE Performance

The purpose of clustering is to identify groups of observations that are cohesive and separate from other groups. We clustered data points into 5 groups using Non-heirarchical K-means clustering via the kmeans() function in the stats package in R using “Weighted RSRP RSRQ Score” and “LTE Signal Power Mean” as variables. Then we used the hierarchical hclust() function with complete-linkage to generate points for visualizing the clusters.