Problem Statement:

TalkingData is China’s largest third-party mobile data platform. Using a SDK that’s integrated with smartphone apps they collect events generated by the smartphone user. This information is used for targeted advertising and mobile analytics.

The objective of this project is to build a model that will predict if a longitude/latitude will be present in a companies log data when a user visits a mobile app. Provided is a given the user’s app usage and smartphone properties.

This could be used to later do behavioral/micro targeting to provide the user with a better advertisement close to their location.

Overview of Technology:

High Level Overview of Steps

1. Software Needed and Installation
2. Configuration Steps
3. Data Description
4. Exploratory Data Analysis
5. Data Transformation
6. Data Modeling

Describe Big Data Set

**URL:**

<https://www.kaggle.com/c/talkingdata-mobile-user-demographics/data>

**Description:**

Data below is mobile platform data from a kaggle competition.   It is very similar to data seen on an advertising exchange.

[TalkingData](https://www.talkingdata.com/), China’s largest third-party mobile data platform, understands that everyday choices and behaviors paint a picture of who we are and what we value. Currently, TalkingData is seeking to leverage behavioral data from more than 70% of the 500 million mobile devices active daily in China to help its clients better understand and interact with their audiences. [1]

**File Sizes:**

* app\_events.csv.zip (211.27 MB)
* App\_labels.csv.zip (4.04 MB)
* Events.csv.zip (62.24 MB)
* Gender\_age\_test.csv.zip (1.05 MB)
* Gender\_age\_train.csv.zip (891.47 KB)
* Label\_categories.csv.zip (7.67 KB)
* Phone\_brand\_device\_model.csv.zip (2.42 MB)

**File Sample Sizes:**

* app\_events.csv.zip ()
* App\_labels.csv.zip ()
* Events.csv.zip ()
* Gender\_age\_test.csv.zip ()
* Gender\_age\_train.csv.zip ()
* Label\_categories.csv.zip ()
* Phone\_brand\_device\_model.csv.zip ()

**Format of data file:**

csv format

Hardware:

* MacBook Pro
* Processor: 2.5 GHz Intel Core 7
* Memory: 16 GB 100 MHz DDR3

Software:

* Juptyer Notebooks (http://jupyter.org/install.html)
* Spark 2.2.1 (https://spark.apache.org/downloads.html)
* PySpark (https://spark.apache.org/docs/0.9.0/python-programming-guide.html)
* PySpark ML (http://spark.apache.org/docs/2.2.0/api/python/pyspark.ml.html)
* Pandas (https://pandas.pydata.org/)
* Numpy (https://www.scipy.org/scipylib/download.html)
* Matplotlib (https://matplotlib.org/users/installing.html)
* Seaborn (https://seaborn.pydata.org/installing.html)

References:

1. <https://www.kaggle.com/c/talkingdata-mobile-user-demographics#description>
2. <https://spark.apache.org/docs/1.1.0/mllib-guide.html>
3. <https://github.com/Azure/Azure-MachineLearning-DataScience/tree/master/Misc/Spark/pySpark>
4. <https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/spark-data-exploration-modeling>

Lessons Learned Pros/Cons

**Lessons Learned:**

Previous to this final I hadn’t really had to do any exploratory data analysis on an unknown dataset.   The final project has taught me to allow multiple days to explore, and get to really know a dataset before going into trying to create a model.  Next time I am testing out features in a model I will start out small.  Started out with way too many features and it was time consuming.

**Pros:**

Got to deep dive into doing a prediction on my own.   I also learned new things about Spark, Pandas, and Machine Learning.   Also, learned how to think about how to assemble data in a format that I could start to extract features for my models

**Cons:**

This project has been super time consuming.   Leaving it up to the student to do anything within the topic ultimately took up a huge chunk of time.   Another big con was choosing to run Spark locally.   Doing my modeling using something like Zepplin/AWS would have really sped things up.

YouTube URLs:

Two Minute URL:

[https://youtu.be/ajY8mxXueoU](https://youtu.be/ajY8mxXueoU" \t "_blank)