~ SENG 474 Project Proposal ~

Alotta ‘clists

by

Devroop Banerjee
Computer Science, Maths
University of Victoria
Victoria, Canada
indroneil20@gmail.com

Yi Xing Hu
Computer Science
University of Victoria
Victoria, Canada
yixinghu@uvic.ca

David Toole
Software Engineering
University of Victoria
Victoria, Canada
static@uvic.ca

Hunter Watson
Software Engineering
University of Victoria
Victoria, Canada
hwatson@uvic.ca

Abstract—**This is a project proposal for the prediction of bikeshare system use, based on the 2011-2012 bikeshare use dataset found on Kaggle.**

I. Introduction

The project has presently just one objective: to forecast bike-sharing rental demand.

Bike-sharing systems provide registered users, as well as non-registered users (referred to as casual users), short-term bike use for a price. Users can rent a bike at one kiosk and return it at a different kiosk, allowing the data to be treated as a sensor network.

The project will use bike-sharing data in combination with date and weather information to predict the demand of bike-sharing rentals within a city. Bike-sharing can be used to extrapolate mobility in a city. This concept can be extended to anticipate the distribution of people across a city based on when and where users are traveling and thereby aid businessmen in opening shops or architects in building houses in profitable locations (the list could go on).

We had also previously hoped to use the data to determine potentially profitable locations to build new kiosks. A visual representation of a city's population density is certainly useful to a lot of businesses. However, unless we find additional data which breaks down demand by kiosk or neighbourhood, this won't be possible.

Technically, this could be considered classification had our program simply sorted locations into 2 classes: -feasible and not feasible. However, since our project doesn't involve results and/or predictions belonging to a certain class, we consider this regressive. A regressive heatmap of the population density of a city would have continuous values spread across the map of city.

II. Related Work

The dataset we've chosen to work with is actually the subject of a Kaggle competition, and thus its study is quite mature. The leaderboard for this competition contains at the time of writing, 3251 entries, representing 3251 attempts to predict bike share use on the basis of weather, season, day of the week, hour of the day, and so on.

A. Predicting bike share use based on weather

Among these was Brandon Harris, who placed 946th, and shared his methodology on his blog.[1]

Aditya Sharma shared his own top fifth-percentile scoring R code through his GitHub account[2], and discussed it on a post on Analytics Vidhya[3].

Giot and Cherrier compared the performance of numerous regressive analysis models in their somewhat exhaustive article on the subject, Predicting bikeshare system usage up to one day ahead[4].

Using these articles, as well as the work on population regression models by Ezequiel Uriel from the University of Valencia[5], we have an excellent basis for our own work in analyzing the data and coming up with an attempt to solve it.

B. Predicting bike share demand geographically

Google has a feature which shows the busiest times of the day when a certain place is searched. Unfortunately, this only shows the busiest times of the day on individual search results (an incomplete picture). We had intended to make a program which essentially made a map of the number of people in different areas of the city based on the drop off and pick up locations of the bikes as the culmination of the second part of our original plan, but have to set aside this goal s it no longer seems feasible.

III. Data Description

Our project will use a dataset by Hadi Fanaee Tork using data from Capital Bikeshare through Kaggle. The data is from Washington, D.C., with a training set collected from the first 19 days of each month. The data fields from Kaggle are listed and summarized in the following table:

Header	Type Information	Description
Datetime	datetime	Hourly date with timestamp
Season	int (1-4)	Integers mapped to the four seasons
Holiday	int (0-1)	Whether the day is a holiday
Workingday	int (0-1)	Whether the day is neither a weekend nor holiday
Weather	int (1-4)	Weather generalization where 1 is clear and 4 is heavy rain.
Temp	float	Temperature in Celsius
Atemp	float	“Feels like” temperature in Celsius
Humidity	int	relative humidity
Windspeed	float	Wind speed
Casual	int	Number of non-registered user rentals initiated
Registered	int	Number of registered user rentals initiated
Count	int	Total number of rentals

Fig. 1. Data fields in Kaggle Bike Sharing Demand dataset

Kaggle also provides a test data set of the remaining days of the month, from the 20th on, of each month for the purpose of scoring, but as the count, casual, and registered fields are missing from this data, we cannot use it for our own purposes. Instead we will train our data on the first 8 of every 10 days in the Kaggle data training set, and validate on the ninth and tenth day of each ten days (training on the first through eighth and eleventh through eighteenth of January 2011, and validating on the ninth, tenth, and nineteenth of January, as well as on the first of February).

IV. Proposed Project

Our project focuses on using the data provided by Kaggle (as mentioned above) as our dataset to train the machine into learning how to extrapolate feasible, densely populated locations. Once the program learns which variables to use in order to make said predictions, it can be provided with datasets similar to the one used, to map out the regressive population density of any city.

V. Timeline

Our proposed timeline for the project is visualized in the following Gantt chart:

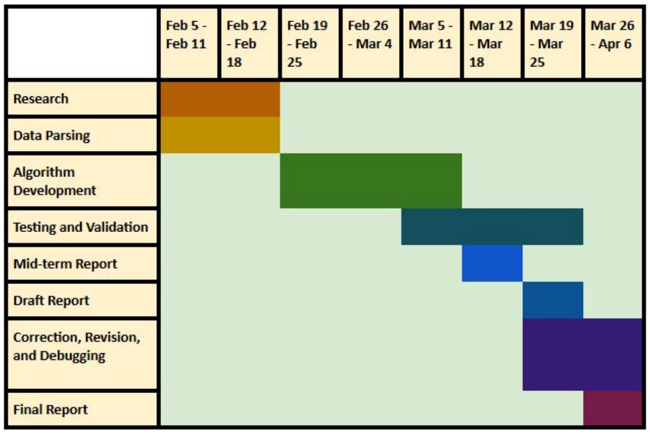

Fig. 2. Proposed timeline for project

VI. Task Distribution

Preliminary breakdown of duties is as follows:

Name	Task
Devroop Banerjee	Leading, Brainstorming, Communication Management
Yi Xing Hu	Data Parsing, Algorithm Development
David Toole	Research, Algorithm Development
Hunter Watson	Document Review, Research, Algorithm Validation and Testing

Fig. 3. Proposed breakdown of tasks by group member

References

[1] B. Harris, “A simple model for Kaggle Bike Sharing,” http://brandonharris.io/kaggle-bike-sharing/, accessed February 6, 2018.

[2] A. Sharma, Bike Sharing Demand, Kaggle, https://github.com/adityashrm21/Bike-Sharing-Demand-Kaggle/blob/master/Bike_Sharing_Demand.R, accessed February 6, 2018.

[3] A. Sharma, “Kaggle Bike Sharing Demand Prediction – How I got in top 5 percentile of participants?,” https://www.analyticsvidhya.com/blog/2015/06/solution-kaggle-competition-bike-sharing-demand/, accessed February 6, 2018

.

[4] R. Giot and R. Cherrier, “Predicting bikeshare system usage up to one day ahead,” published in Computational Intelligence in Vehicles and Transportation Systems (CIVTS), 2014 IEEE Symposium on, http://ieeexplore.ieee.org/document/7009473/, accessed February 6, 2018.

[5] E. Uriel, “The simple regression model: estimation and properties,” https://www.uv.es/uriel/2%20Simple%20regression%20model%20estimation%20and%20properties.pdf, accessed February 6, 2018