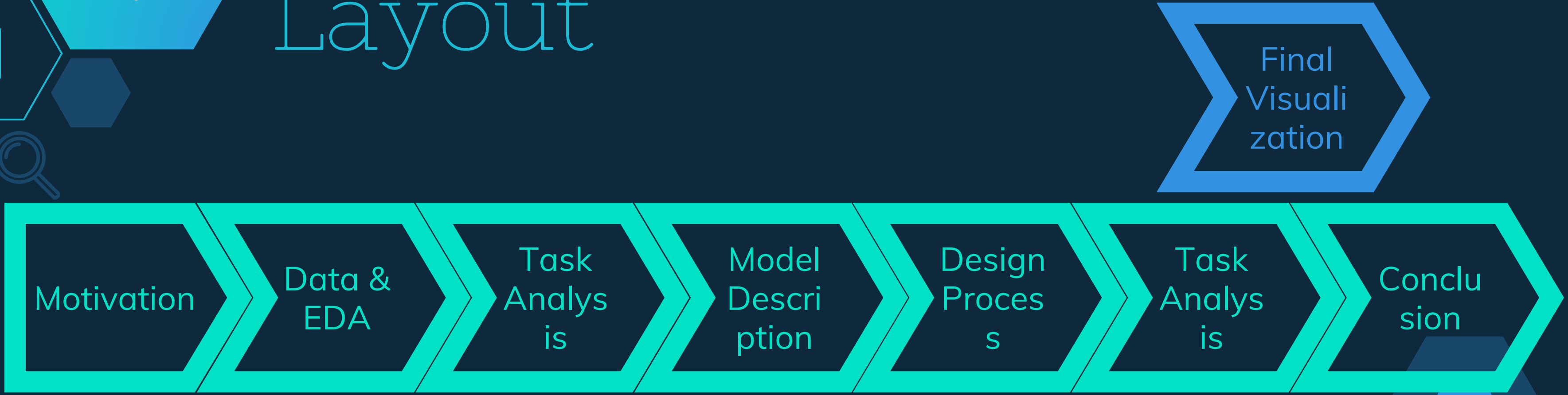
City Recommendation in the USA using Yahoo Flickr Creative Commons 100M Dataset

by Anant Jain, Ahmet Salih Gündoğdu

DS 5500 Fall 2018 --- Prof. Cody Dunne, Northeastern University

Layout



Motivation

Data & EDA

Task Analysis

Model Description

Design Process

Task Analysis

Conclusion

Final Visualization

1

Motivation

Let’s start with the first set of slides

facebook
The Batcave
ROBIN
just checked in
You are still the mayor
The location of our Bat Cave is meant to be secret, so STOP CHECKING IN!

A picture is worth a thousand words

A complex idea can be conveyed with just a single still image, namely making it possible to absorb large amounts of data quickly.

2 Data & EDA

Yahoo Flickr Creative Commons 100M

In short, YFCC100M

- One of the largest assemblages of multimedia check-ins ever created
- Publicly hosted on AWS
- Released under the Yahoo Web-Scope program
- Hundred million media objects dating between 2004 and 2014

Yahoo Flickr Creative Commons (YFCC100M)

Pruning

ELIMINATED UNWANTED COLUMNS

- Workable with limited RAM
- Omitting records that weren't geo-tagged (i.e. more than 50%)
- Omitting records that came with a wrong date format (0.01%)

FILTER TO USA

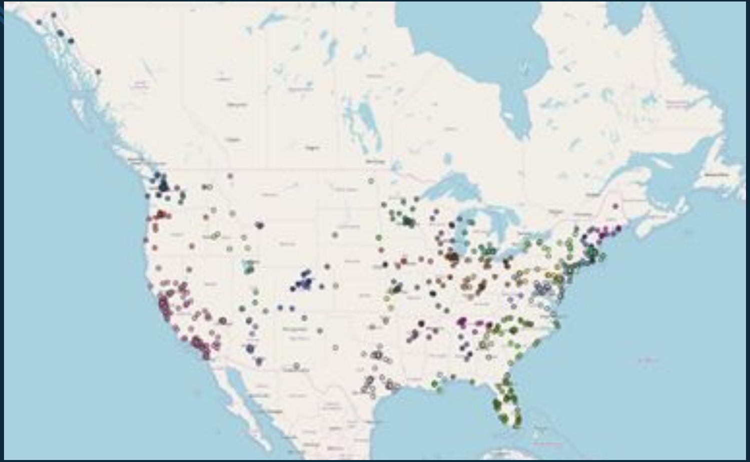
- YFCC100M Places - Expansion Dataset
- Reverse geocode information of all records.

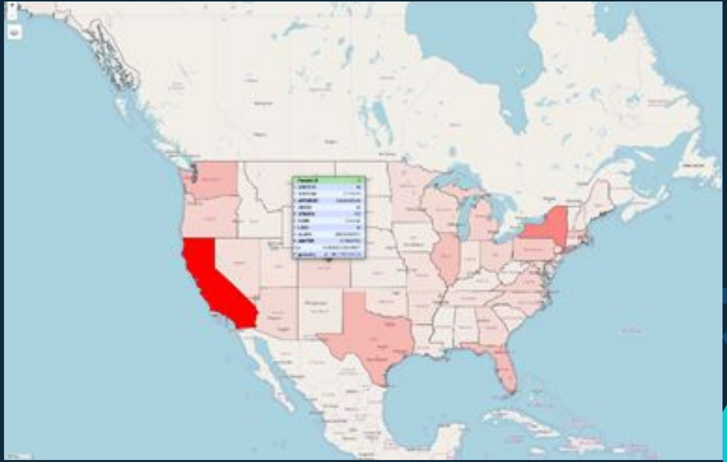
YFCC100M + Pruning + Merging + Cleaning = YFCC_USA16M

Columns

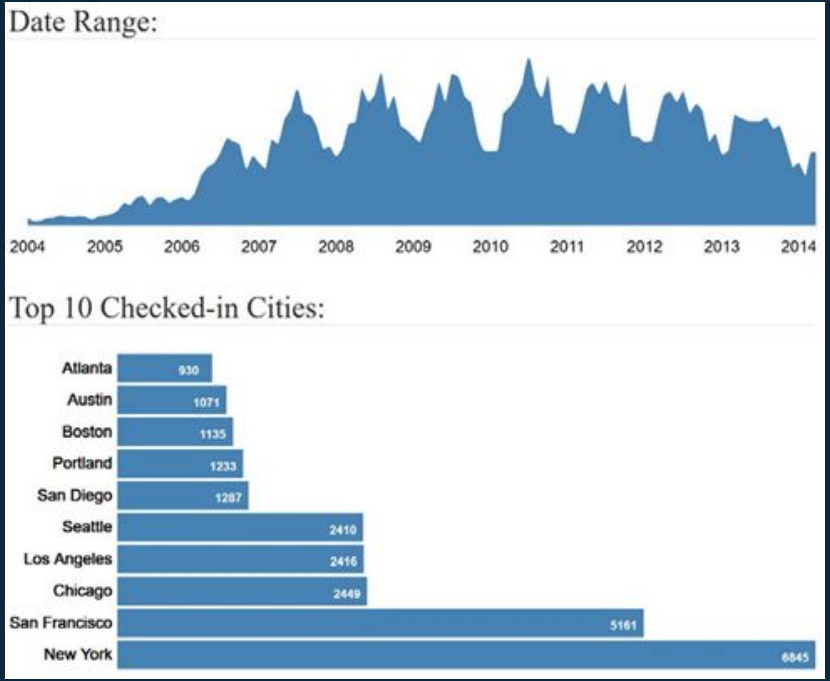
pid	Unique media identifier
user_nickname	User identifier
date	Date the media object was created
longitude	Longitude of the location the media object was checked at
latitude	Latitude of the location the media object was checked at
town	Town the media object was checked in
state	State the media object was checked in

EDA





EDA (contd.)



Date Range:

Top 10 Checked-in Cities:

Objective

Utilize the travel check-in data and use data-based visualizations to explore, assess and evaluate multiple SVD algorithms for the purposes of identifying anomalies, generating trust and providing the best recommendation for cities to visit in the USA

3

Task Analysis

Tasks

Priority	Domain Task	Analytic Task	Search Task	Analyze Task
3	Examining and evaluating the model performance of the recommended places against the given user's travel history	Compare	Locate	Present
2	Generate a ranked list of recommendations	Sort	Explore	Present
1	Visualize different models and hyperparameters for assessment of the best set of modeling parameters to use.	Compare	Explore	Discover
4	Exploratory Data Analysis	Compare	Explore	Discover

Intended Users

Experts

Researchers and machine learning engineers who are interested in recommendation systems.

Travelers

Anybody who wants to get travel recommendations in the USA

4

Model Description

Backend

Assorted selection of Hyperparameters and Models

PREPROCESSING

- Numeric: #
- Binary: 1 or 0

MODELS

- SVD_explicit
- SVD_implicit: Alternating Least Squares

LATENT DIMENSIONS

- Number of dimensions/features to extract for each user and location

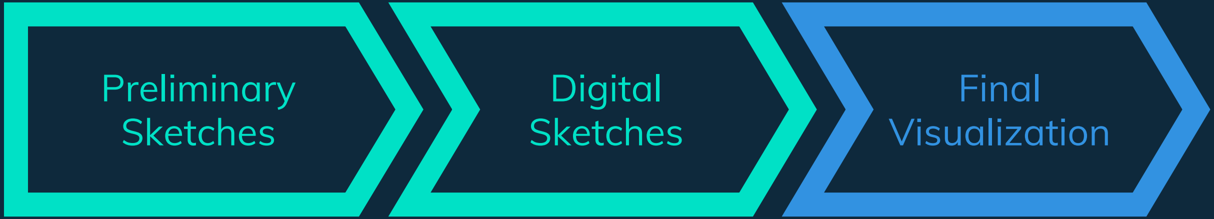
METRIC

- Precision-Train Set
- Recall-Train Set
- Precision-Validation Set
- Recall-Validation Set

5

Design Process

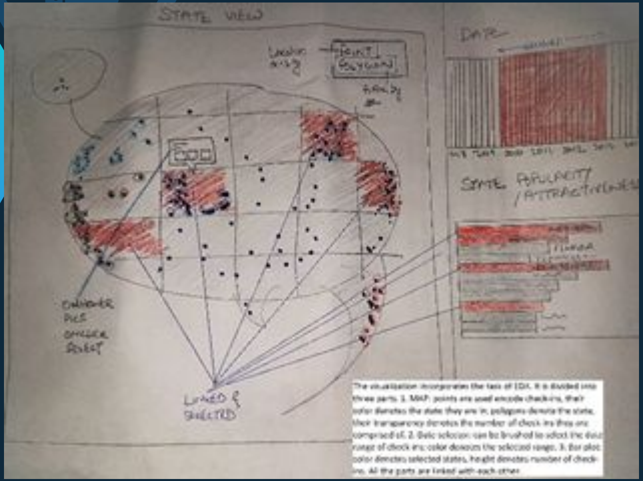
Design Process



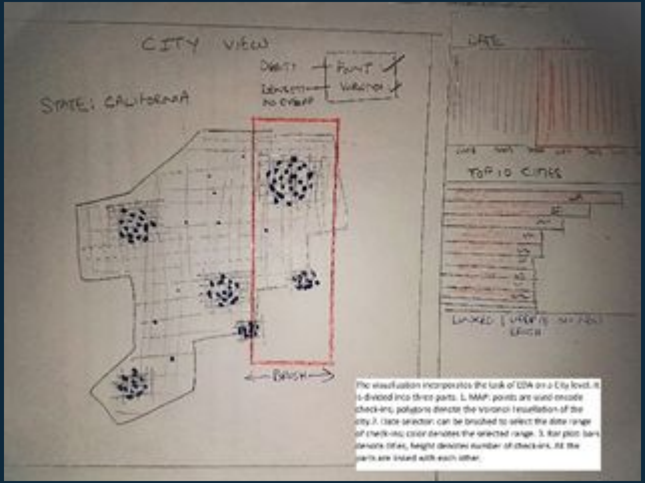
Preliminary Sketches

Digital Sketches

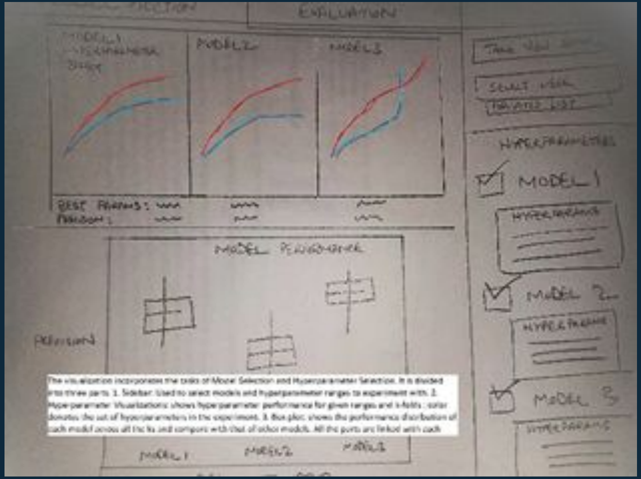
Final Visualization



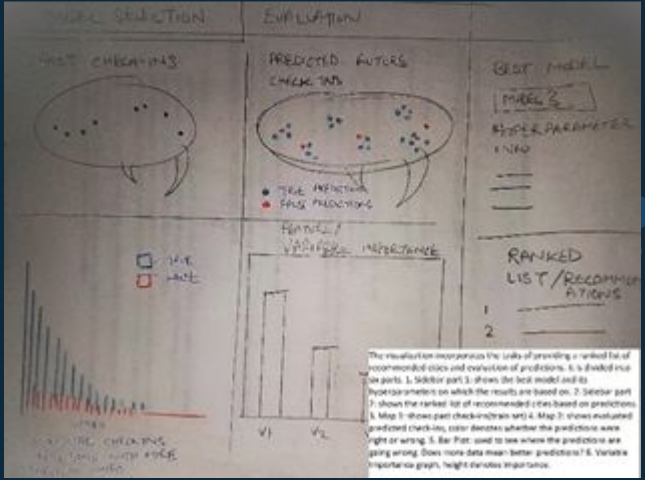
The visualization incorporates the task of EDA. It is divided into three parts. 1. MAP: points are used encode check-ins, their color denotes the state they are in, polygons denote the state, their transparency denotes the number of check ins they are comprised of. 2. Date selection: can be brushed to select the date range of check ins; color denotes the selected range. 3. Bar plot: color denotes selected states, height denotes number of check-ins. All the parts are linked with each other.



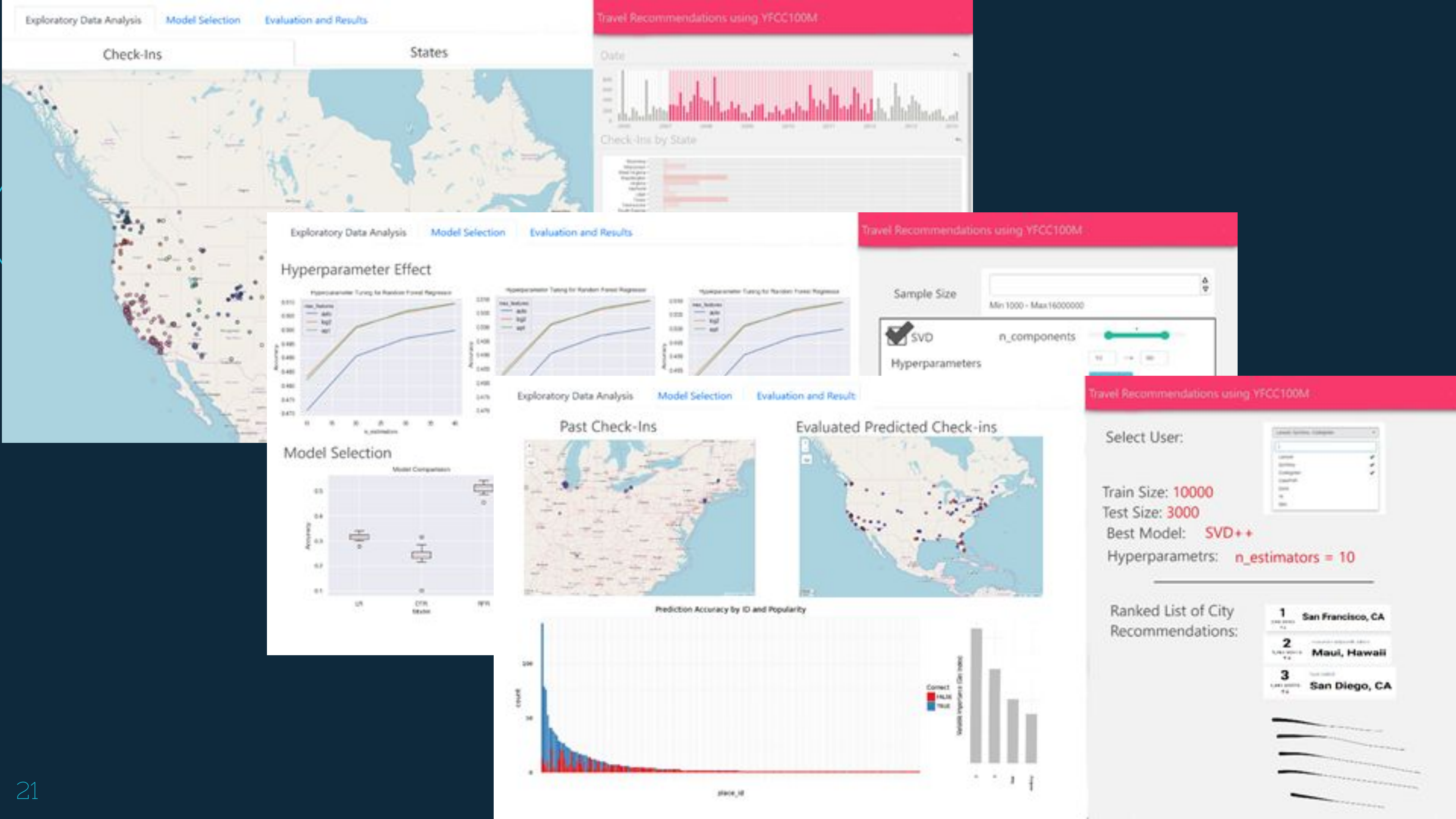
The visualization incorporates the task of EDA on a City level. It is divided into three parts. 1. MAP: points are used encode check-ins, polygons denote the voronoi tessellation of the city 2. Date selection: can be brushed to select the date range of check-ins; color denotes the selected range. 3. Bar plot: bars denote cities, height denotes number of check-ins. All the parts are linked with each other.



The visualization incorporates the tasks of Model Selection and Hyperparameter Selection. It is divided into three parts. 1. Sidebar: Used to select models and hyperparameter ranges to experiment with. 2. Hyperparameter Visualizations: shows hyperparameter performance for given ranges and k-folds ; color denotes the set of hyperparameters in the experiment. 3. Box plot: shows the performance distribution of each model across all the ks and compare with that of other models. All the parts are linked with each



The visualization incorporates the tasks of providing a ranked list of recommended cities and evaluation of predictions. It is divided into six parts. 1. Sidebar part 1: shows the best model and its hyperparameters on which the results are based on. 2. Sidebar part 2: shows the ranked list of recommended cities based on predictions. 3. Map 1: shows past check-ins(train set) 4. Map 2: shows evaluated predicted check-ins, color denotes whether the predictions were right or wrong. 5. Bar Plot: used to see where the predictions are going wrong. Does more data mean better predictions? 6. Variable importance graph, height denotes importance.



Exploratory Data Analysis
Model Selection
Evaluation and Result
Past Check-Ins
Evaluated Predicted Check-ins
Prediction Accuracy by ID and Popularity
count
place_id
Correct
FALSE
TRUE
Travel Recommendations using YFCC100M

Select User:

Train Size: 10000

Test Size: 3000

Best Model: SVD++

Hyperparametrs: n_estimators = 10

Ranked List of City Recommendations:

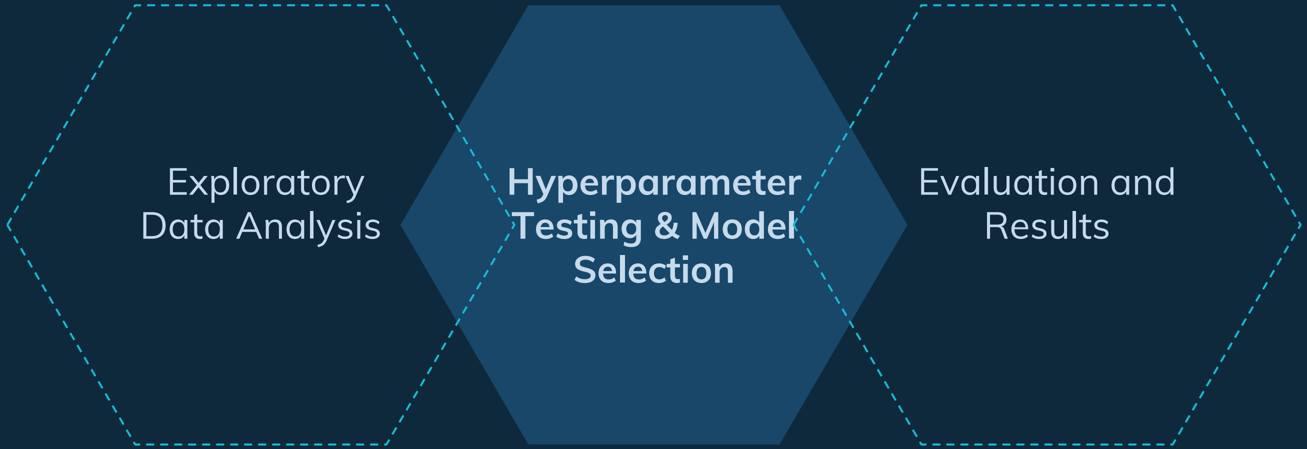
1 San Francisco, CA

2 Maui, Hawaii

3 San Diego, CA

Video Walkthrough

Final Visualization

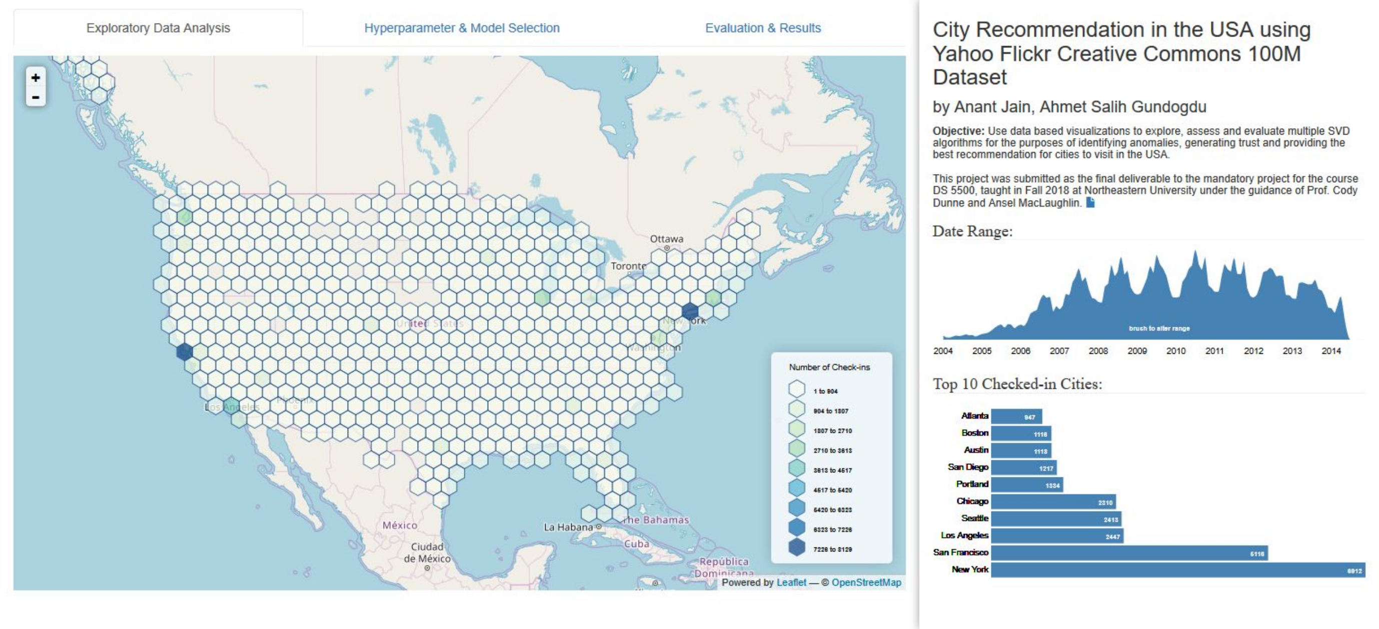


Exploratory Data Analysis

Hyperparameter Testing & Model Selection

Evaluation and Results

Exploratory Data Analysis



City Recommendation in the USA using Yahoo Flickr Creative Commons 100M Dataset

by Anant Jain, Ahmet Salih Gundogdu

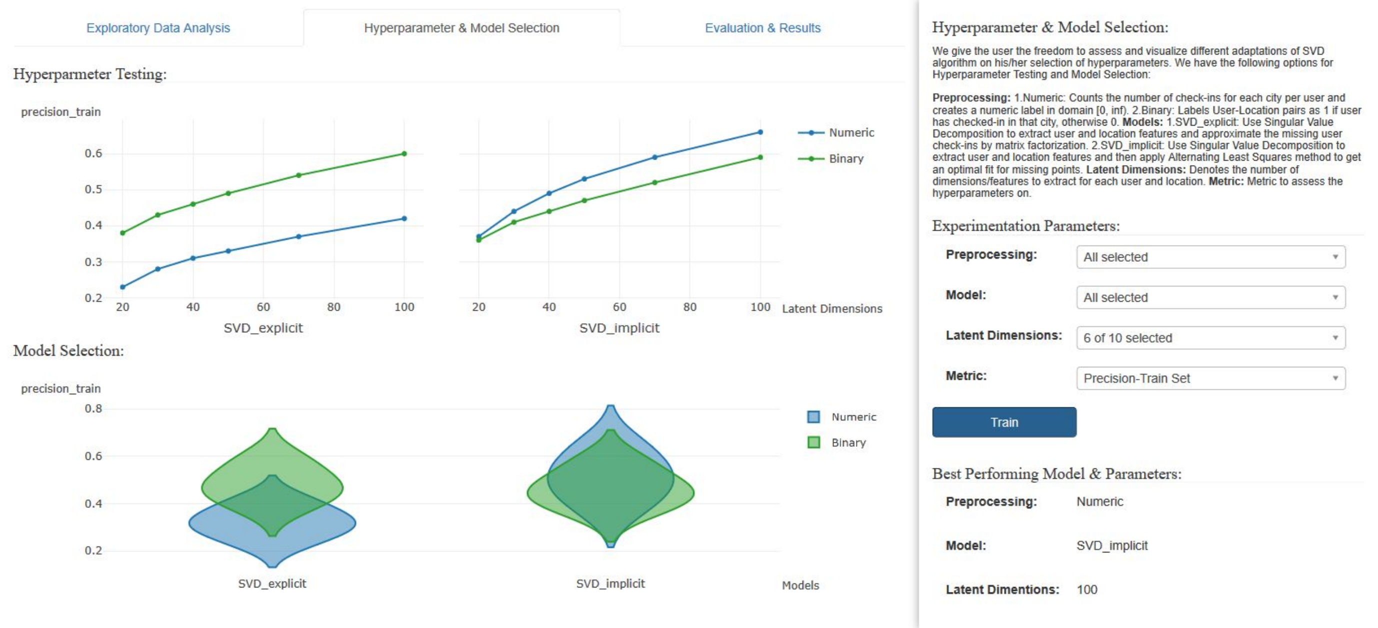
Objective: Use data based visualizations to explore, assess and evaluate multiple SVD algorithms for the purposes of identifying anomalies, generating trust and providing the best recommendation for cities to visit in the USA.

This project was submitted as the final deliverable to the mandatory project for the course DS 5500, taught in Fall 2018 at Northeastern University under the guidance of Prof. Cody Dunne and Ansel MacLaughlin.

Date Range:

Top 10 Checked-in Cities:

Hyperparameter Testing & Model Selection



Hyperparmeter Testing:

Model Selection:

Hyperparameter & Model Selection:

We give the user the freedom to assess and visualize different adaptations of SVD algorithm on his/her selection of hyperparameters. We have the following options for Hyperparameter Testing and Model Selection:

Preprocessing: 1.Numeric: Counts the number of check-ins for each city per user and creates a numeric label in domain [0, inf). 2.Binary: Labels User-Location pairs as 1 if user has checked-in in that city, otherwise 0. **Models:** 1.SVD_explicit: Use Singular Value Decomposition to extract user and location features and approximate the missing user check-ins by matrix factorization. 2.SVD_implicit: Use Singular Value Decomposition to extract user and location features and then apply Alternating Least Squares method to get an optimal fit for missing points. **Latent Dimensions:** Denotes the number of dimensions/features to extract for each user and location. **Metric:** Metric to assess the hyperparameters on.

Experimentation Parameters:

Preprocessing: All selected

Model: All selected

Latent Dimensions: 6 of 10 selected

Metric: Precision-Train Set

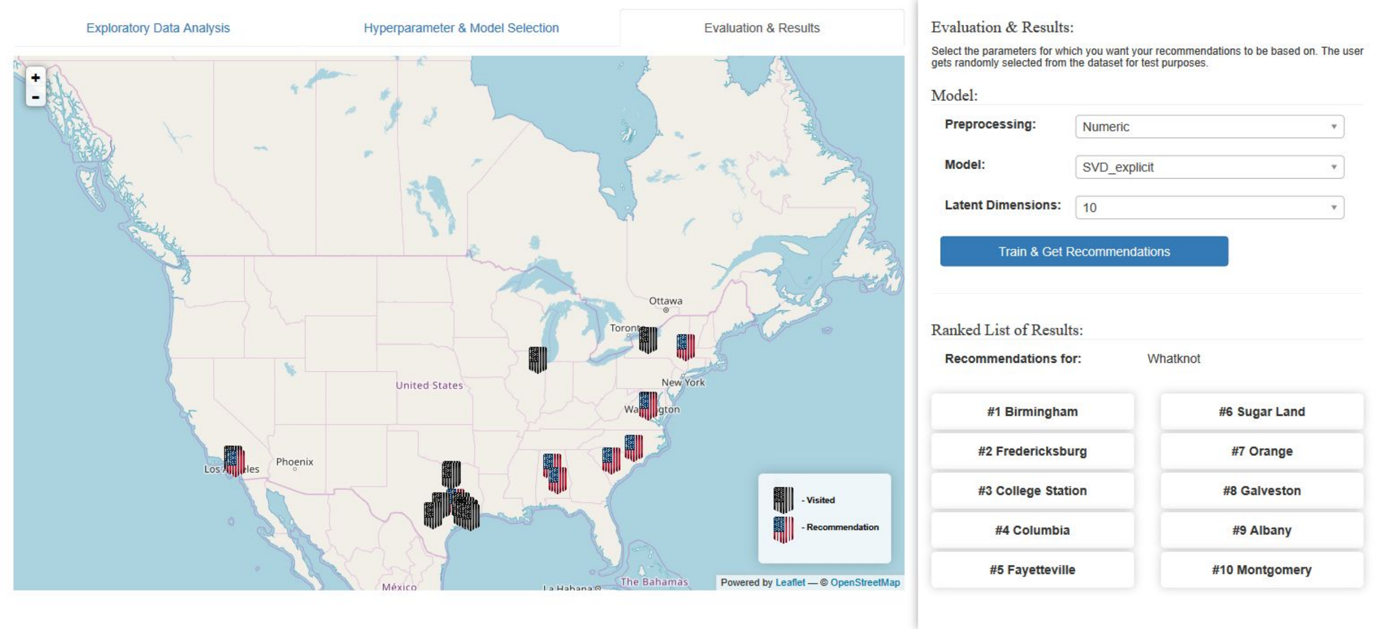
Best Performing Model & Parameters:

Preprocessing: Numeric

Model: SVD_implicit

Latent Dimentions: 100

Evaluation and Results



Evaluation & Results:

Select the parameters for which you want your recommendations to be based on. The user gets randomly selected from the dataset for test purposes.

Model:

Preprocessing: Numeric

Model: SVD_explicit

Latent Dimensions: 10

Ranked List of Results:

Recommendations for: Whatknot

#1 Birmingham

#2 Fredericksburg

#3 College Station

#4 Columbia

#5 Fayetteville

#6 Sugar Land

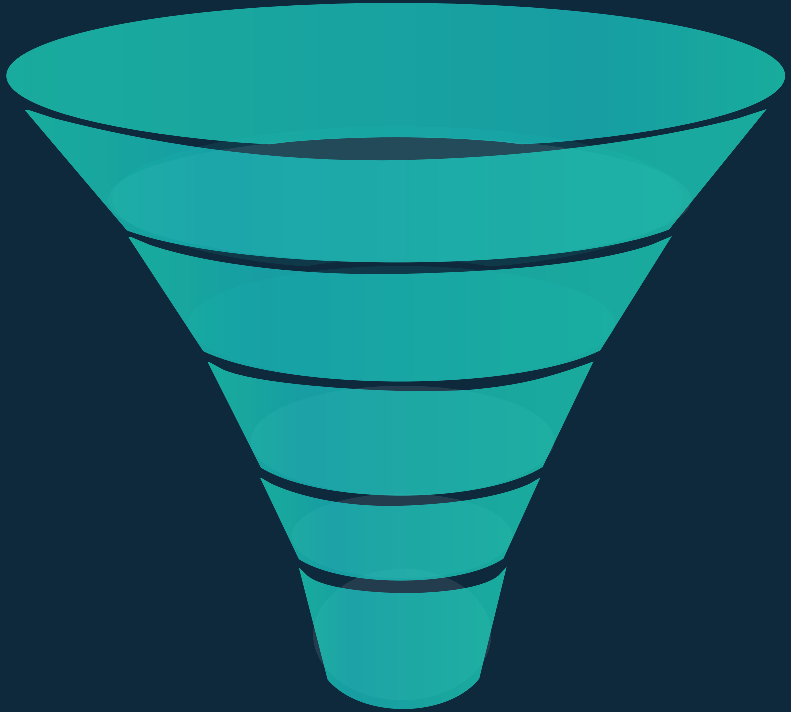
#7 Orange

#8 Galveston

#9 Albany

#10 Montgomery

Conclusion



Bind ML with Visualizations

Proper Visual Encodings

Include User in the ML tasks

Build Trust in Results

Enjoyment :)

Future Work

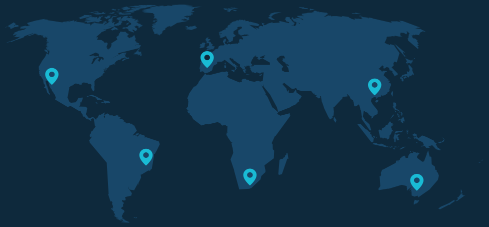
Integration of more complex models

E.x. Autoencoders

Better evaluation techniques

Distance-based, etc.

Scale to cover the whole world instead of just the US



Thanks!

Any questions?

You can find us at:

- https://github.com/antujn
- https://github.com/asgundogdu

Github URLs are attached to the icons.

Credits

Special thanks to all the people who made and released these awesome resources for free:

- d3
- leaflet
- colorbrewer
- tipsy
- plotly
- flask
- bootstrap
- multi-select
- pylab
- implicit