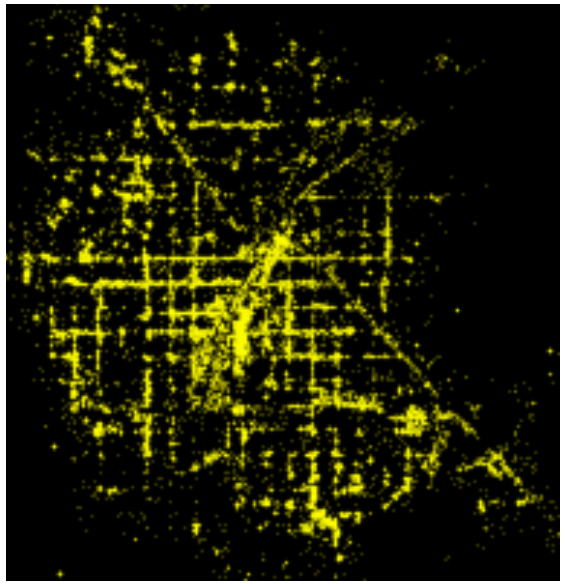
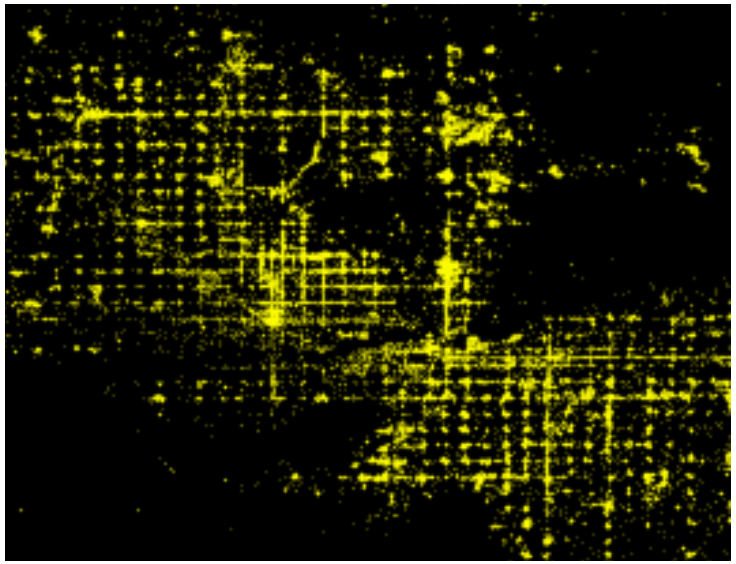
Utilizing Yelp Businesses Categories and price ranges to predict neighborhoods affluency



Authors: Jacopo Cecchi, Jason Leung



AGENDA

I. OBJECTIVES AND METHODOLOGY

II. EXECUTIVE SUMMARY

III. DATA ACQUISITION AND PREPROCESSING

IV. MODELLING & EVALUATION

V. APPLICATION DEVELOPMENT

VI. CONCLUSIONS

AGENDA

I. OBJECTIVES AND METHODOLOGY

II. EXECUTIVE SUMMARY

III. DATA ACQUISITION AND PREPROCESSING

IV. MODELLING & EVALUATION

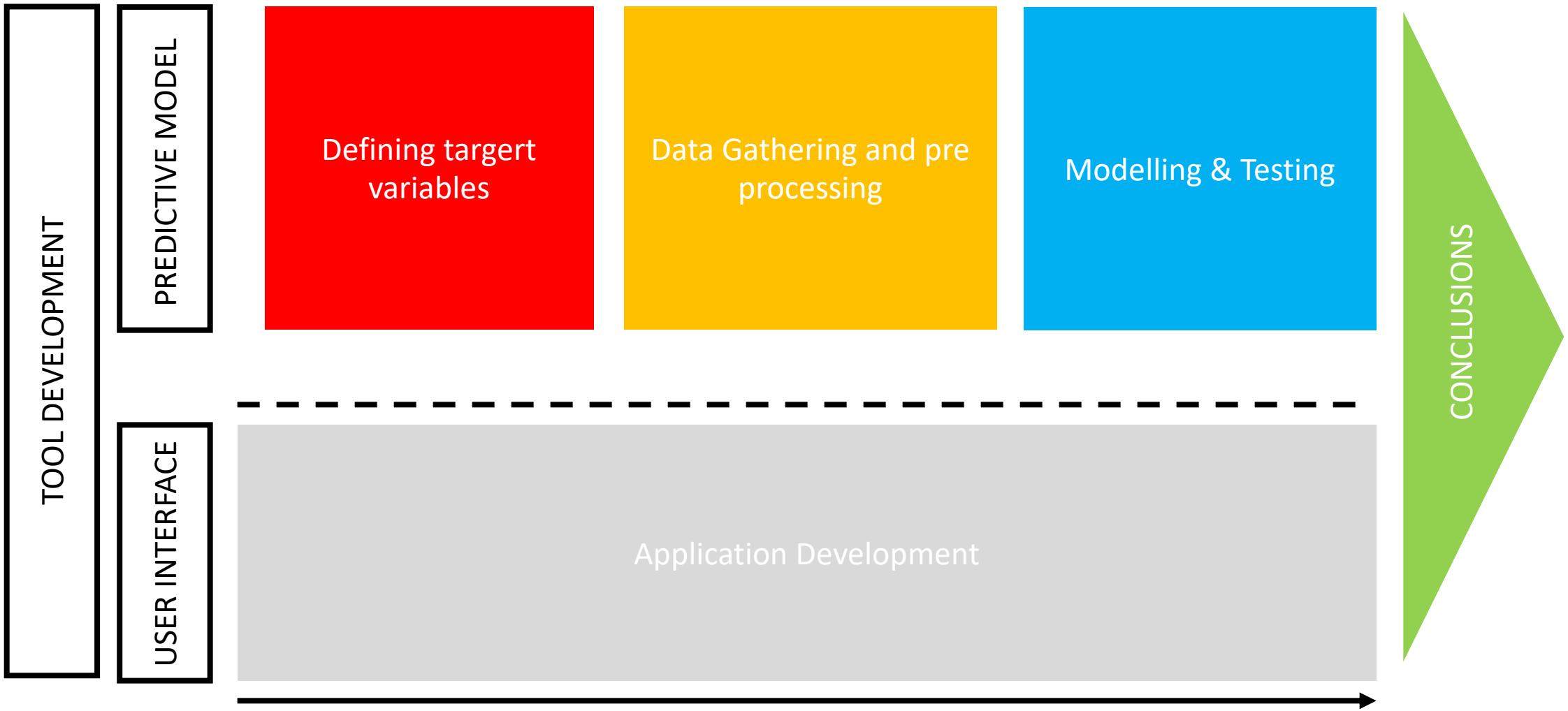
V. APPLICATION DEVELOPMENT

VI. CONCLUSIONS

DEVELOP A TOOL TO ESTIMATE NEIGHBORHOOD AFFLUENCY LEVERAGING YELP PRICES ESTIMATES

Novelty of the approach: use of big data related to commercial activity and cost of products and services as indicator of affluency.

METHODOLOGY



AGENDA

I. OBJECTIVES AND METHODOLOGY

II. EXECUTIVE SUMMARY

III. DATA ACQUISITION AND PREPROCESSING

IV. MODELLING & EVALUATION

V. APPLICATION DEVELOPMENT

VI. CONCLUSIONS

EXECUTIVE SUMMARY 1/2

- A working tool that allows to retrieve affluency estimates for a given area given Yelp's businesses and services costs estimates has been developed and is available for deployment.
- Several supervised learning models – Logistic Regression, KNN, CART, Random Forest, Bagging Classifier, Adaboost and SVC have been trained and tested on a total count of 7 US cities. Among these, KNN resulted to be the best performing model.
- While models perform well – F1 score of ca. 90% on test sets of cities on which they are trained, their performance is inaccurate on testing sets of cities on which they have not been trained. We believe that this outcome is partially due to the fact that algorithms manage to identify geographic patterns in the data even if they are not fed businesses geographic coordinates directly.

EXECUTIVE SUMMARY 2/2

- Our tool is relatively reliable at estimating a city's neighbours affluency, BUT only when it has been trained on a set of observation of the specific city. Without a city-specific training, the model is not good at serving its purpose as it is accurate ca. less than 20% of the time.
- This study shows possibilities in predicting per capita income of a locality based just on Yelp estimates of surrounding businesses and services activities and prices.
- Further research could explore if models trained on a wide range of US cities would give better baseline predictions of affluency in non-training US cities.

AGENDA

I. OBJECTIVES AND METHODOLOGY

II. EXECUTIVE SUMMARY

III. DATA ACQUISITION AND PREPROCESSING

IV. MODELLING & EVALUATION

V. APPLICATION DEVELOPMENT

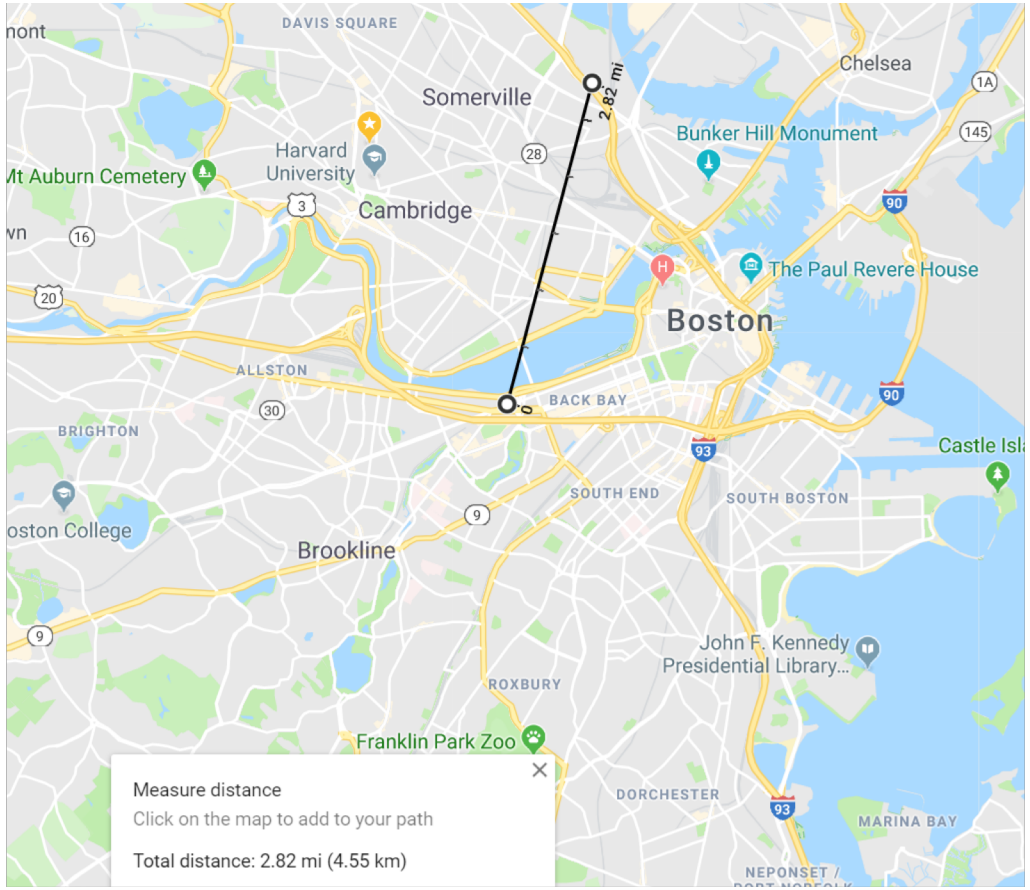
VI. CONCLUSIONS

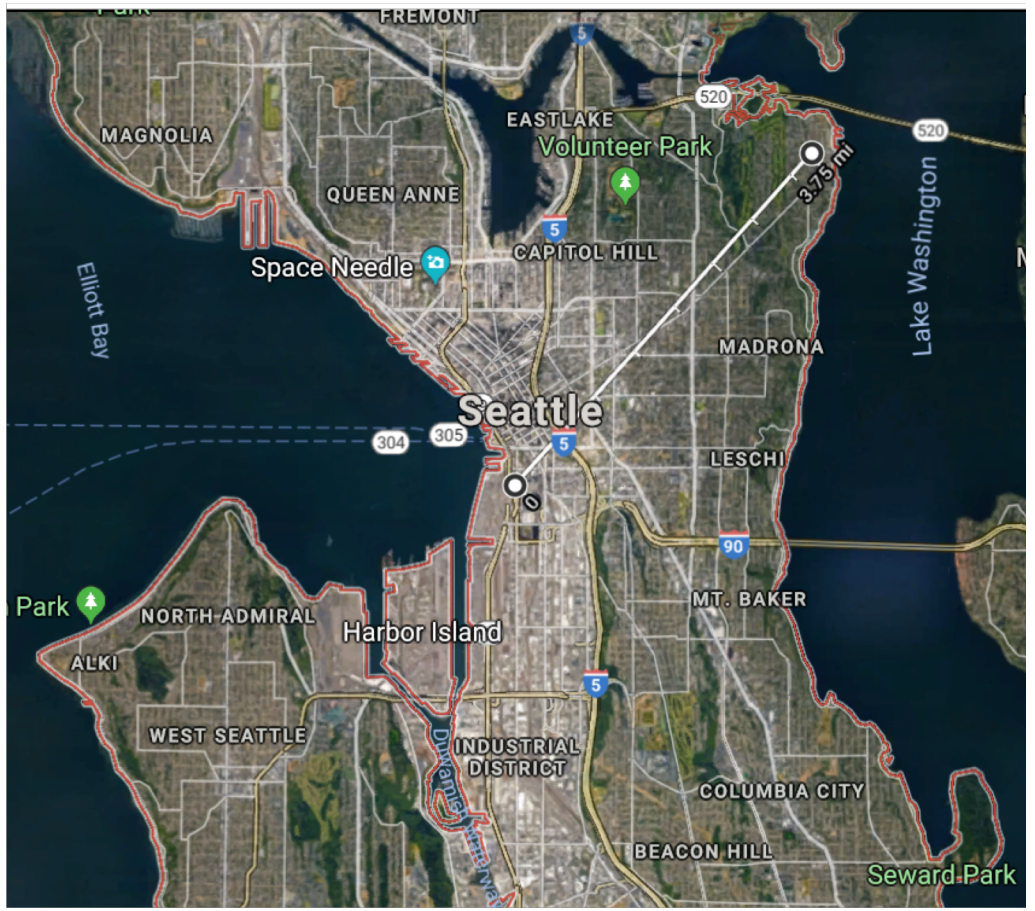
WE SELECTED AV. INCOME PER CAPITA AS TARGET VARIABLE – DECLINED ALONG INCOME BRACKETS

- Av. Income per capita has been selected arbitrarily as proxy for a locality affluency and as target variable for our models.
- Models have been trained to classify instances across income brackets – see table on the right.
- When models scores are reported, models refer to scores on 9 classes unless it is specified otherwise.
- Income data sourced from American Community Survey (ACS) 2017 Individual Income by Block Group

Bracket Range 9 Classes	Bracketing 1	Bracketing 2	Bracketing 3
$0-$10.99k	1	1	1
$11-$20.99k	2	2	1
$21-$30.99k	3	3	2
$31-$40.99k	4	4	2
$41-$50.99k	5	5	3
$51-$60.99k	6	6	3
$61-$70.99k	7	7	4
$71-$80.99k	8	8	4
$81+k.99	9	8	4

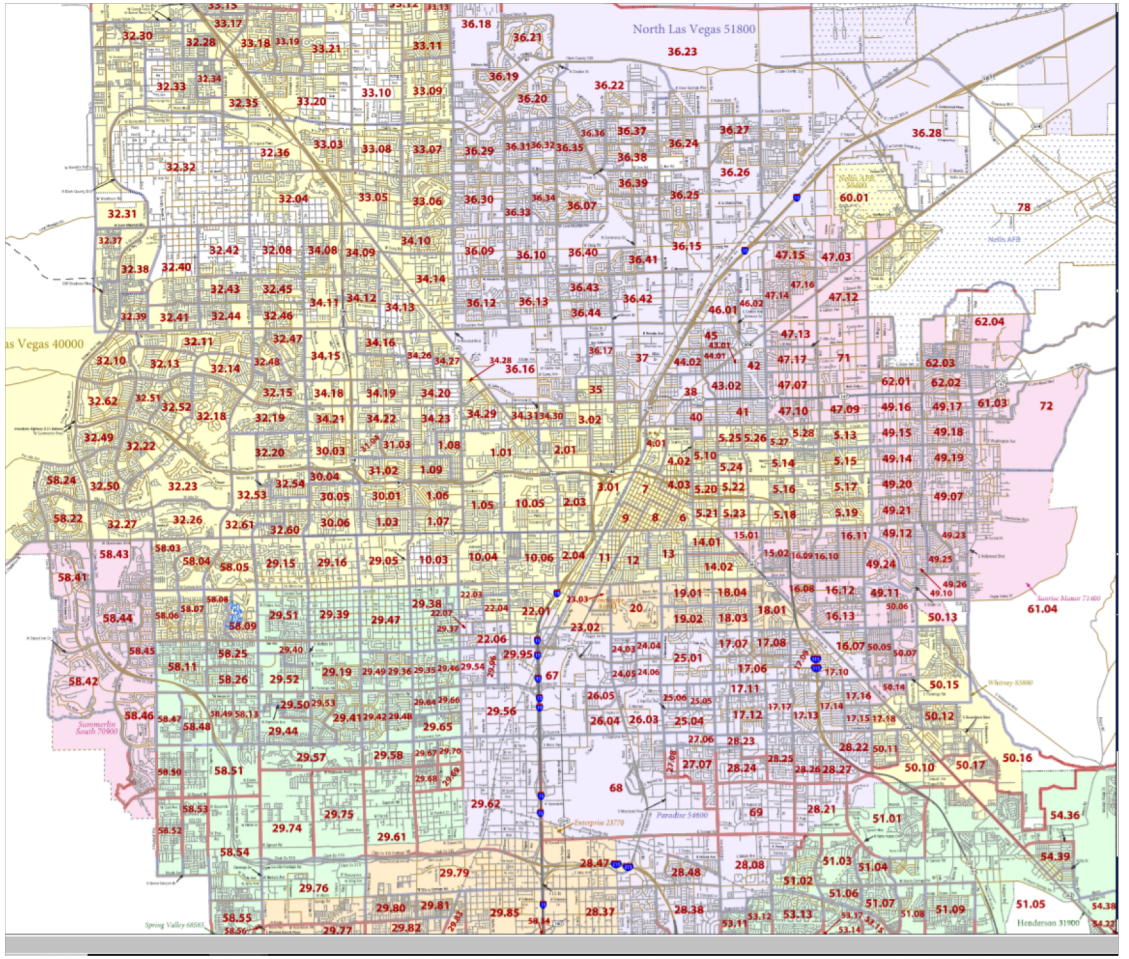
WHICH CITIES, WHAT RADIUS? YELP RESTRICTIONS AND CITY PARTICULARS INFLUENCED OUR CHOICES



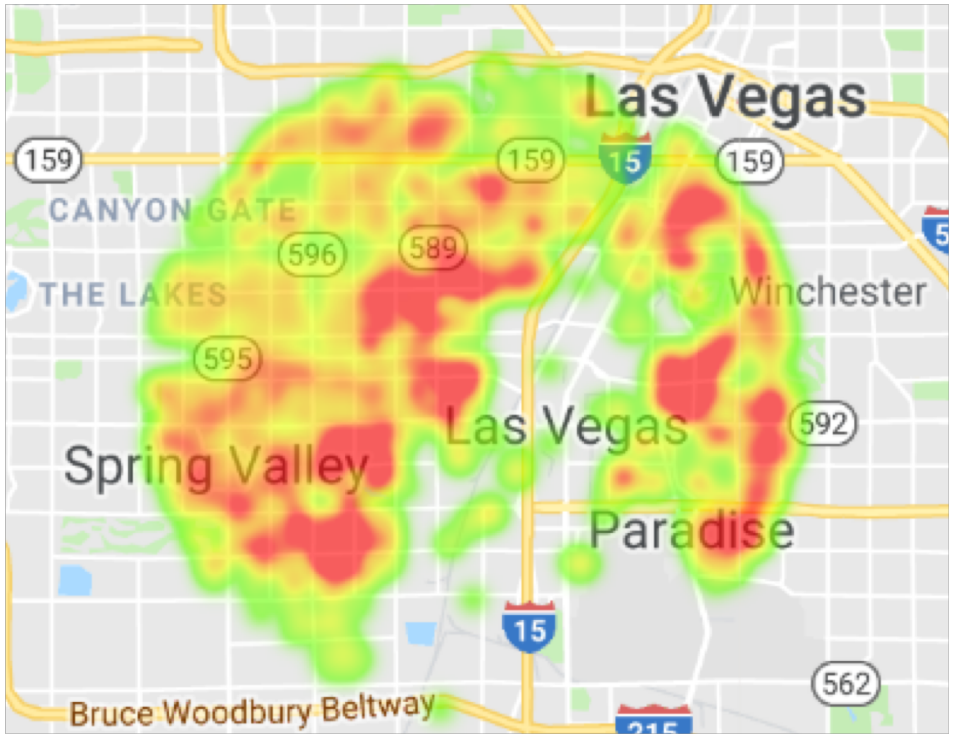


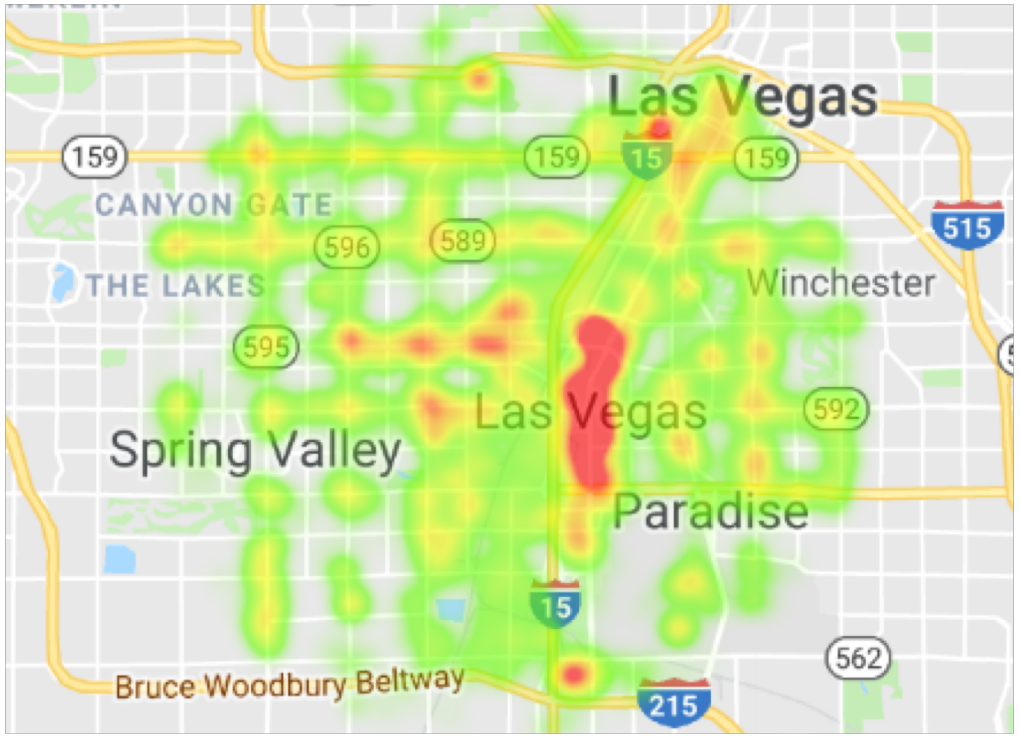
ORGANIZING GEOGRAPHIC DATA: ZIPCODE INCOMES VERSUS CENSUS TRACT AND BLOCK GROUP GRANULARITY

- Census Data grouped by Tract (see image on right)
- 2017 Income from American Community Survey by Block Group, one level down
- Bottom Census level is by block, ~50-500 population from 2010 Decennial Census
- Exact Lat/Longitudes of Every Block also from Census
- Dataset Structure: For every 100 people per block, find all business types and price range within 1000 meter radius of the block centroid
- Are there predictive patterns between a block's average income and the business pricing and categories around it?



Uncovering Patterns Between Population Density and Business Density within City Centers





WE LEVERAGED CA. 24K OBSERVATIONS FROM 7 CITIES

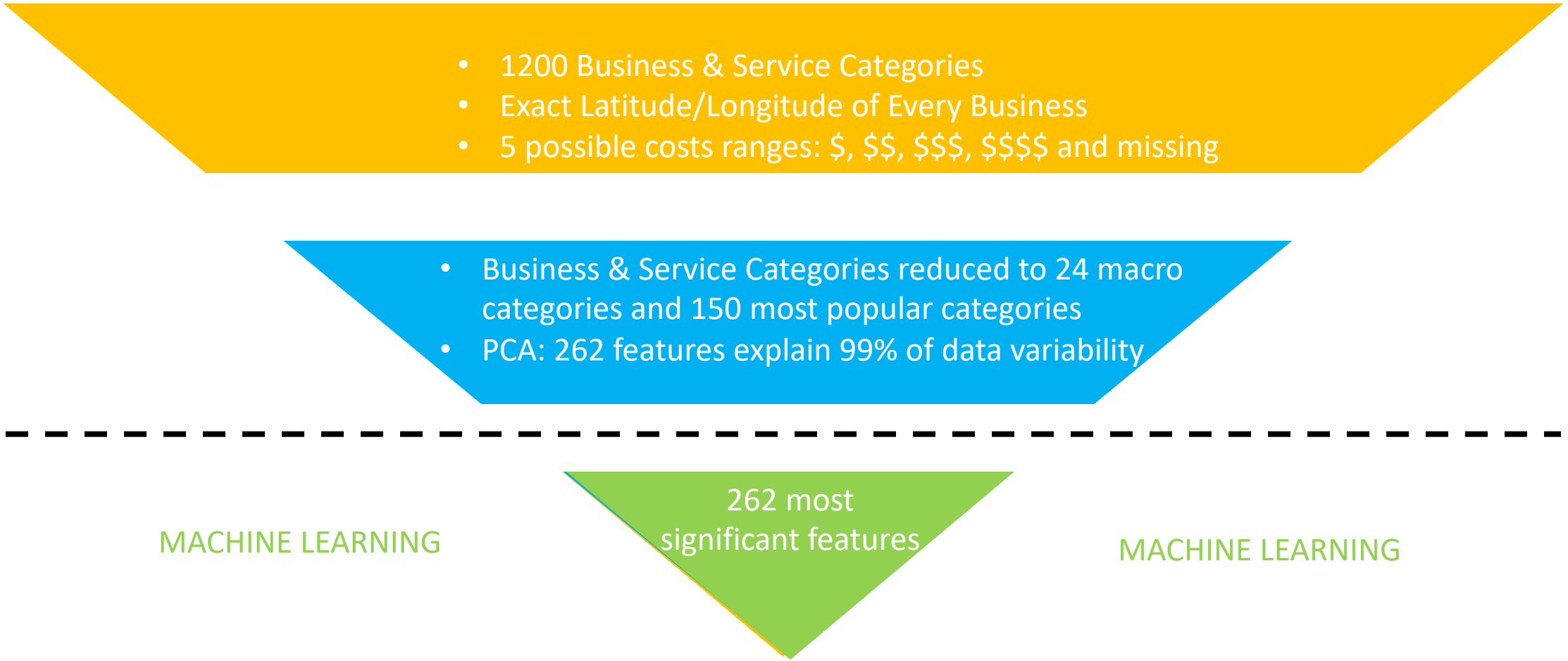
City	Total Population (2016)	Population Density
Phoenix, AZ	1,1615,017	3,126 (517 sq.mi)
Charlotte, NC	842,051	2,929 (298 sq.mi)
Las Vegas, NV	632,912	4,660 (136 sq.mi.)
Pittsburgh, PA	303,625	5,484 (55 sq.mi.)
Seattle, WA	704,352	8,391 (84 sq.mi)
Boston, MA	673,184	13,943 (61 sq.mi)
Miami, FL	453,579	12,645 (36 sq.mi)
Dallas, TX* (comparably not enough businesses)	1,317,929	3,870 (341 sq.mi)

THROUGH YELP'S API AND THROUGH KAGGLE'S DATASET

KAGGLE DATASET (years through 2013)
Charlotte, NC
Pittsburgh, PA
Las Vegas, NV
Phoenix, AZ

YELP API (years through 2019)
Seattle, WA
Boston, MA
Miami, FL
Las Vegas, NV
Phoenix, AZ
Charlotte, NC

TOO MANY FEATURES LED US TO PERFORM FEATURES REDUCTION



AGENDA

I. OBJECTIVES AND METHODOLOGY

II. EXECUTIVE SUMMARY

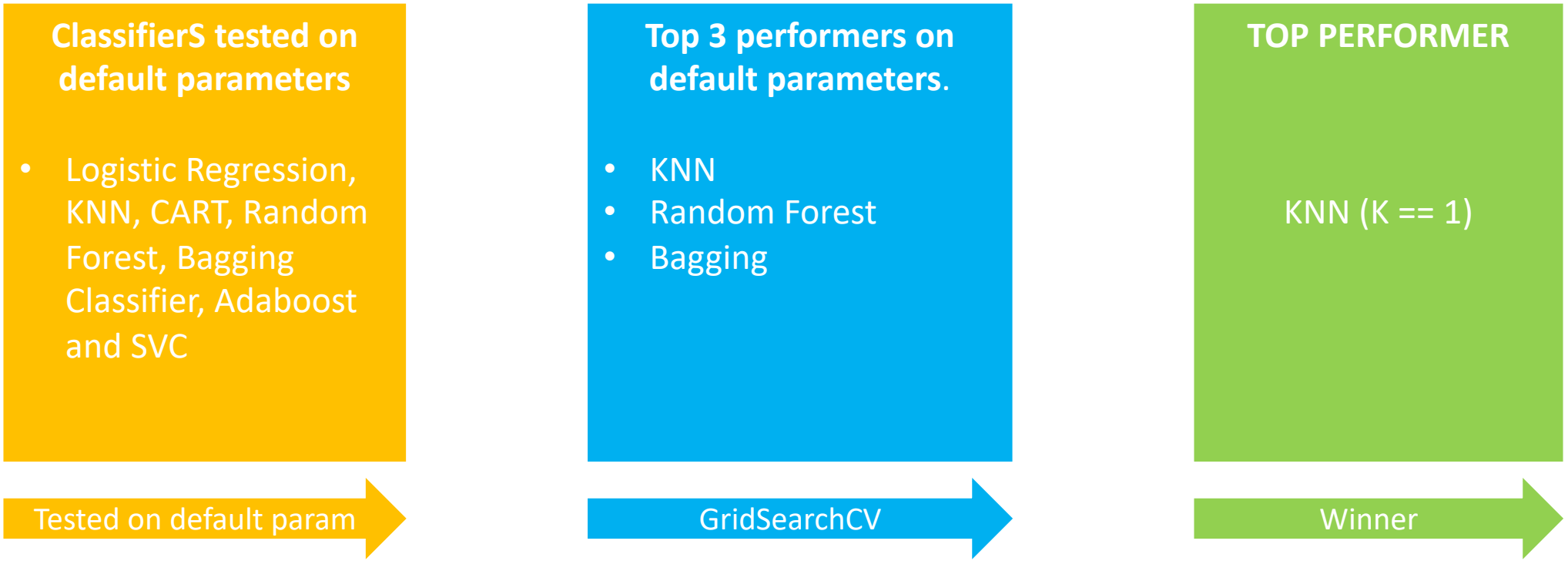
III. DATA ACQUISITION AND PREPROCESSING

IV. MODELLING & EVALUATION

V. APPLICATION DEVELOPMENT

VI. CONCLUSIONS

KNN WITH K==1 OUTPERFORMED THE OTHER 6 CLASSIFIERS CONSIDERED FOR THE EXPERIMENT



MODELS PERFORM WELL WHEN TESTED ON THE SAME CITIES ON WHICH HAVE BEEN TRAINED (1/2)

Tests results of models trained and tested on same 7 cities, 9 income classes.

Model	Training F1 score	Testing F1 score
KNeighborsClassifier	94%	82%
RandomForestClassifier	93%	81%
BaggingClassifier	96%	84%

MODELS PERFORM WELL WHEN TESTED ON THE SAME CITIES ON WHICH HAVE BEEN TRAINED (2/2)

Tests results of models trained and tested on same 7 cities, 9 income classes.

Model	Training F1 score	Testing F1 score
KNeighborsClassifier	98%	88%
RandomForestClassifier	92%	82%
BaggingClassifier	98%	87%

…BUT POORLY WHEN TESTED ON CITIES THEY HAVE NOT BEEN TRAINED ON

Tests results of models trained on 4 cities and tested on 3 different cities, 9 income classes.

Model	Training F1 score	Testing F1 score
KNeighborsClassifier	16%	18%
RandomForestClassifier	19%	20%
BaggingClassifier	20%	21%

ALL-AROUND MEDIOCRE SCORES WERE OBTAINED ON UNTRAINED CITIES, ESPECIALLY ON TYPES OF CITIERS DIFFERENT TRAINING TYPES.

Trained on Las Vegas, Phoenix, Pittsburgh and Charlotte, tested on combined set of Boston, Miami, Seattle

Income Variable Reduced to 4 Brackets: [0-20k, 20-40k, 40-60k, 60+k]

with observation counts: (7410, 3998, 1148, 943)

Model	Accuracy	Testing F1 score
Kneighbors	39%	30%
Logistic Regression Classifier	36%	56%
RandomForestClassifier	38%	8%
BaggingClassifier	39%	17%

“BEST” MEDIOCRE SCORE FOR MODELS ON UNTRAINED CITIES: 6 CITY TRAINED LOGISTIC REGRESSION ON LAST CITY (MIAMI) (1/2)

Income Variable Reduced to 3 Brackets: 0-20k, 20-50k, 50+k (better class balance)

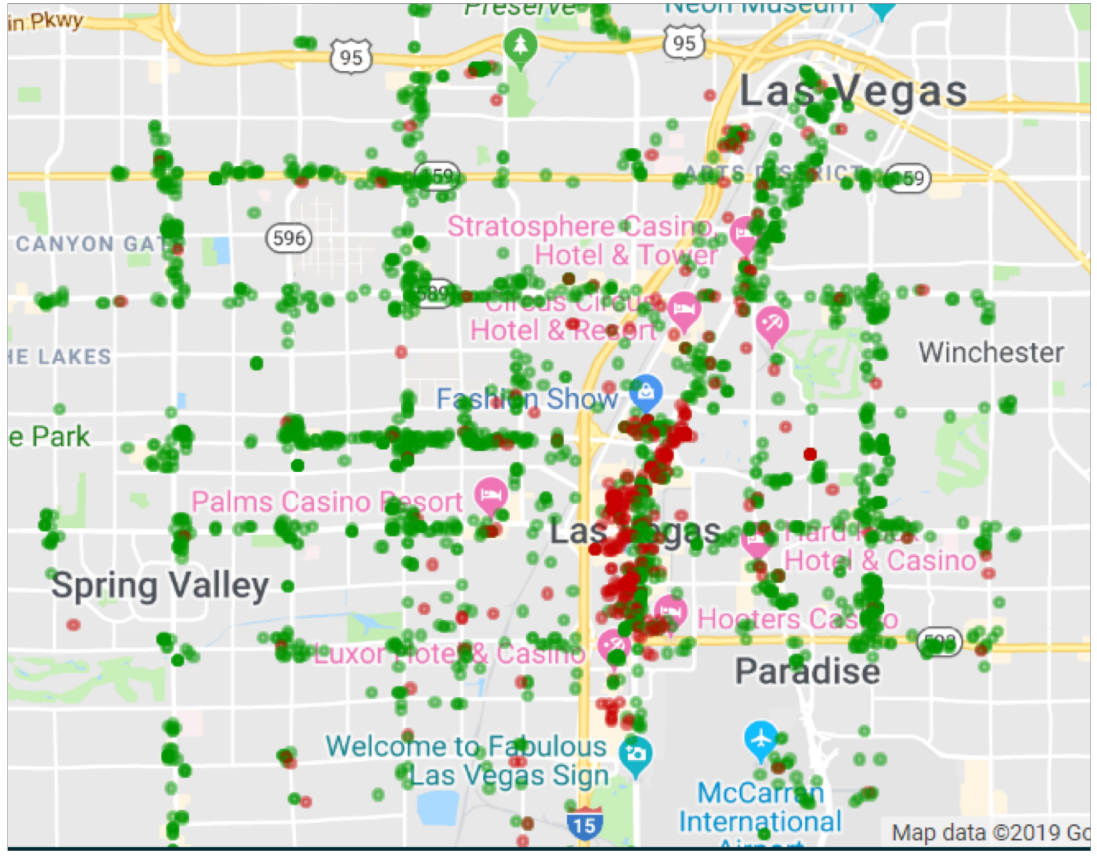
Model	Accuracy	Testing F1 score $0-$20k	Testing F1 score $20-$50k	Testing F1 score $50k+
Kneighbors	41%	55%	21%	14%
Logistic Regression Classifier	46%	63%	18%	13%
RandomForestClassifier	39%	51%	29%	8%
BaggingClassifier	41%	53%	31%	10%

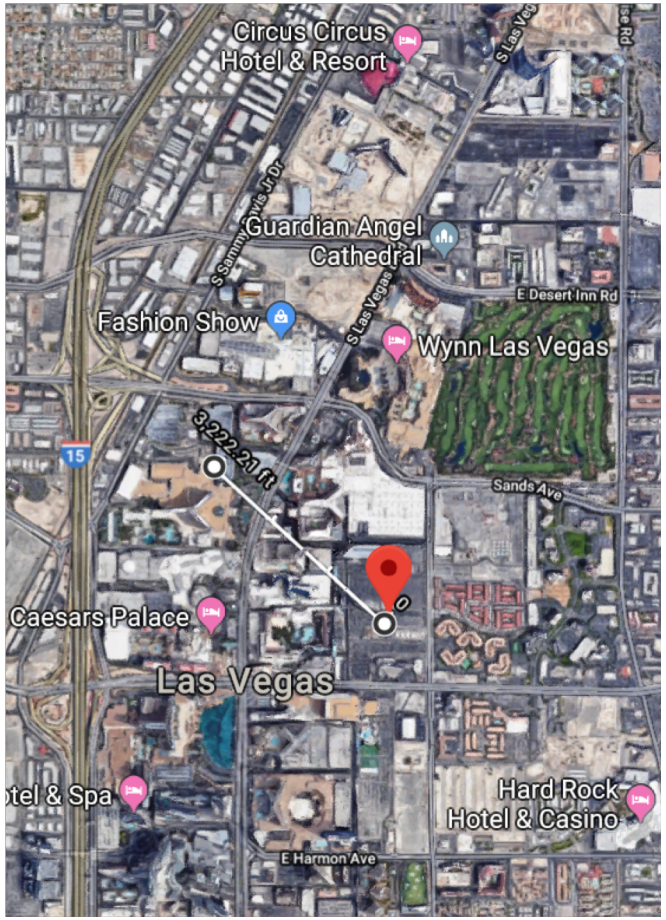
“BEST” MEDIOCRE SCORE FOR MODELS ON UNTRAINED CITIES: 6 CITY TRAINED LOGISTIC REGRESSION ON LAST CITY (MIAMI) (2/2)

Income Variable Reduced to 3 Brackets: 0-20k, 20-50k, 50+k (better class balance)

Confusion Matrix	Predicted Low Income	Predicted Mid Income	Predicted High Income
Actual Low Income	41%	21%	9%
Actual Mid Income	16%	5%	7%
Actual High Income	1%	1%	2%

HIGH PRICE BUSINESS CLUSTERS CORRELATE WITH SOME HIGH AFFLUENCY LOCALES BUT ONLY FOR SOME CITIES





AGENDA

I. OBJECTIVES AND METHODOLOGY

II. EXECUTIVE SUMMARY

III. DATA ACQUISITION AND PREPROCESSING

IV. MODELLING & EVALUATION

V. APPLICATION DEVELOPMENT

VI. CONCLUSIONS

Application on GitHub:

run through GoogleMaps_UserSearch.ipynb

AGENDA

I. OBJECTIVES AND METHODOLOGY

II. EXECUTIVE SUMMARY

III. DATA ACQUISITION AND PREPROCESSING

IV. MODELLING & EVALUATION

V. APPLICATION DEVELOPMENT

VI. CONCLUSIONS

CONCLUSIONS 1/2

- A working tool that allows to retrieve affluency estimates for a given area given Yelp's businesses and services costs estimates has been developed and is available for deployment.
- Several supervised learning models – Logistic Regression, KNN, CART, Random Forest, Bagging Classifier, Adaboost and SVC have been trained and tested on a total count of 7 US cities. Among these, KNN resulted to be the best performing model.
- While models perform well – F1 score of ca. 90% on test sets of cities on which they are trained, their performance is inaccurate on testing sets of cities on which they have not been trained. We believe that this outcome is partially due to the fact that algorithms manage to identify income clusters in the data even if they are not fed geographic coordinates directly.

CONCLUSIONS 2/2

- Our tool is relatively reliable at estimating a city's neighbours affluency, BUT only when it has been trained on a set of observations of the specific city. Without a city-specific training, the model is not good at serving its purpose as it is accurate ca. less than 20% of the time.
- This study shows possibilities in predicting per capita income of a locality based just on Yelp estimates of surrounding businesses and services activities and prices.
- Further research could explore if models trained on a wide range of US cities would give better baseline predictions of affluency in non-training US cities.