nyc neighborhood affluence:

yelp edition

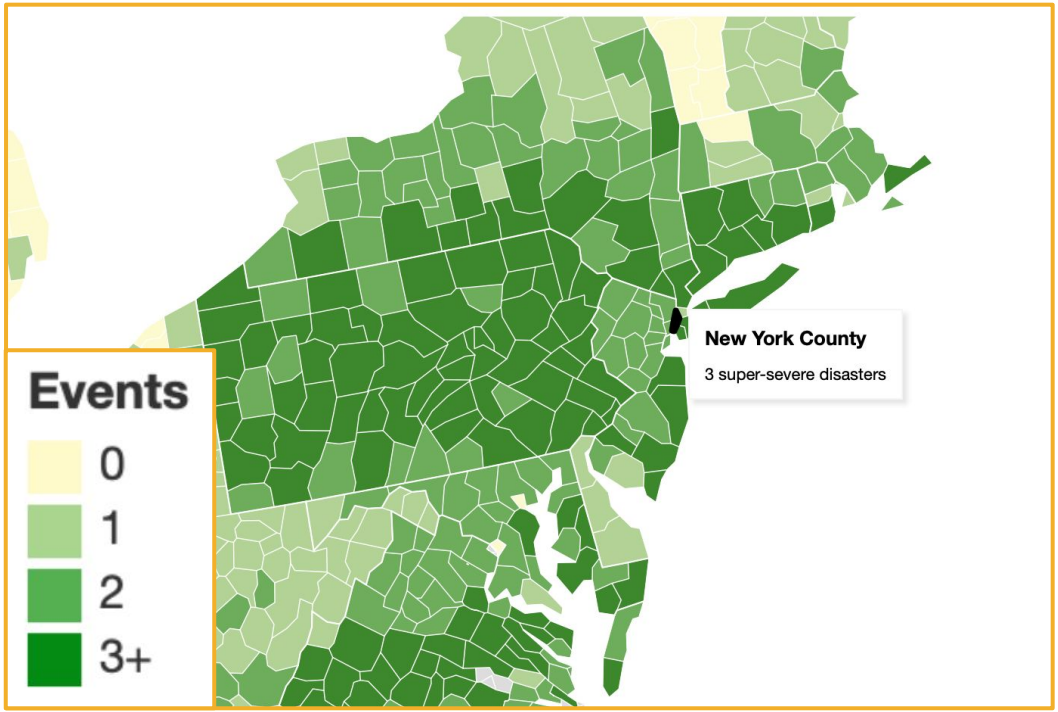
by: eddie yip, hadi morrow, mahdi shadkam-farrokhi

Problem Statement

- **Utilizing Yelp “$” price to estimate neighborhood affluency**
- **What is Yelp “$” price?**
- **How can we help?**

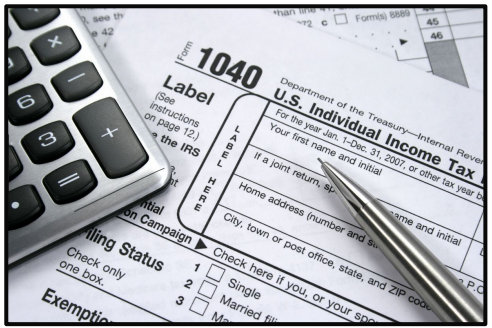
Problem Statement

- **Why NYC?**
- **1930 - 2010**



Data Collection

- **Yelp Fusion API**
  - Yelp “$” Price
  - Review Counts
  - Zip Code
- **IRS Data**
  - 2016 tax return data from irs.gov



Cleaning Data

Yelp

- Yelp “$” prices with missing values
- Businesses located outside of NYC

IRS

- Excel file
- Missing some zip codes!
  - Using these to test model!

EDA

- Creating Affluency rate
  - Why 15%?

Zip Code 10001

Total IRS Returns = 50,000
$200k+ Returns = 10,000

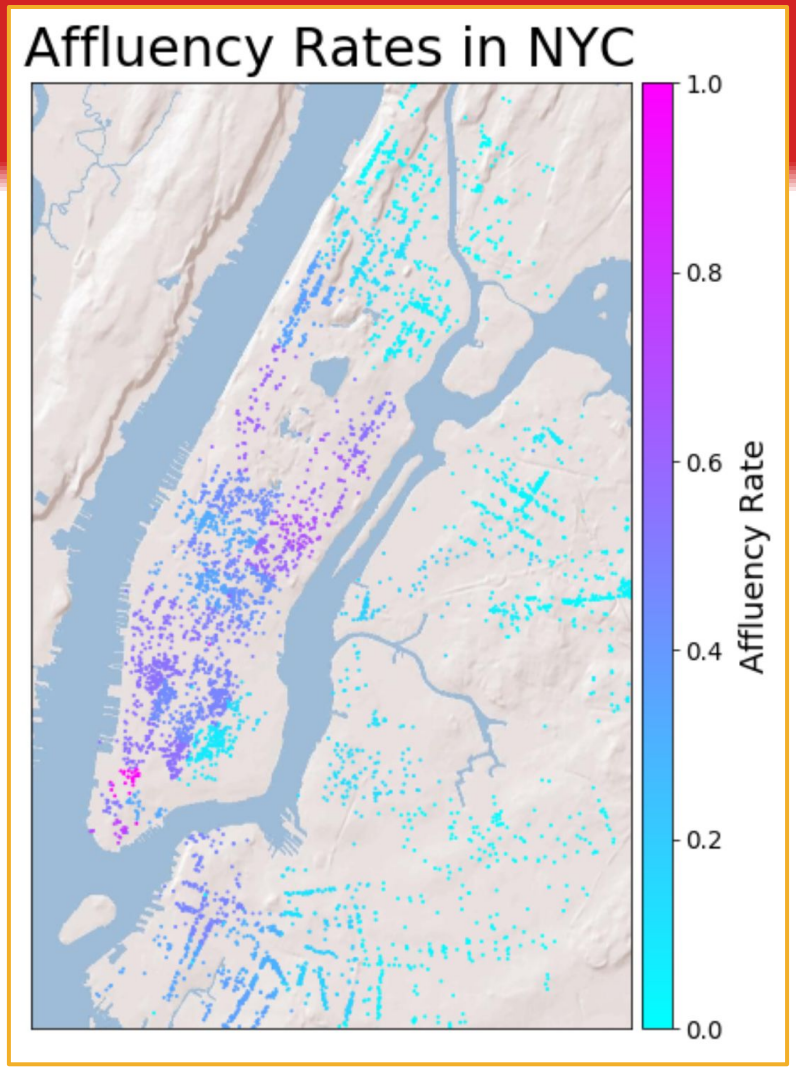
Affluency rate = 20%
Is affluent? TRUE

10001 New York, NY

EDA

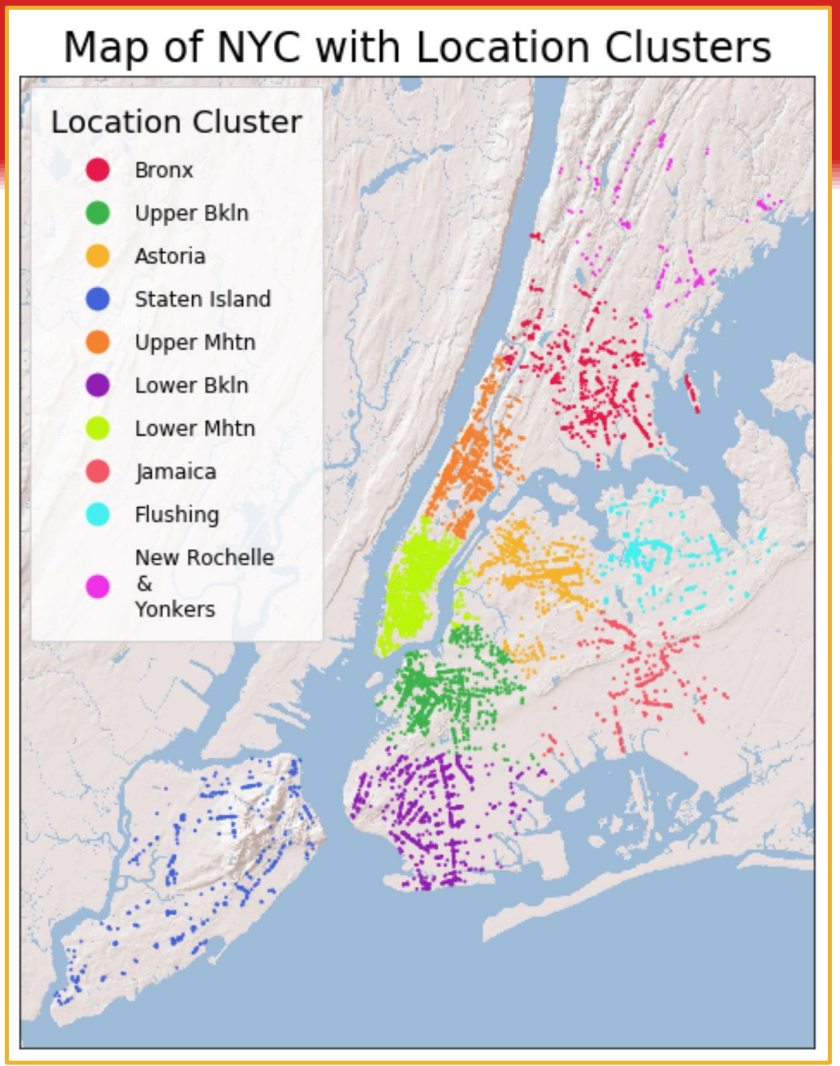
- High Affluence in Manhattan

- How does our data capture this connection between location and affluence?



Feature Engineering

- Location clustering using K-means
- Clear distinction b/t location clusters
- One-Hot Encoding
  - Location clusters
  - Categories, transactions
- Price * Rating Score



Model Preparation

- **Features**
  - Categories & transactions
  - Location clusters
  - Price
  - Rating
  - Review count

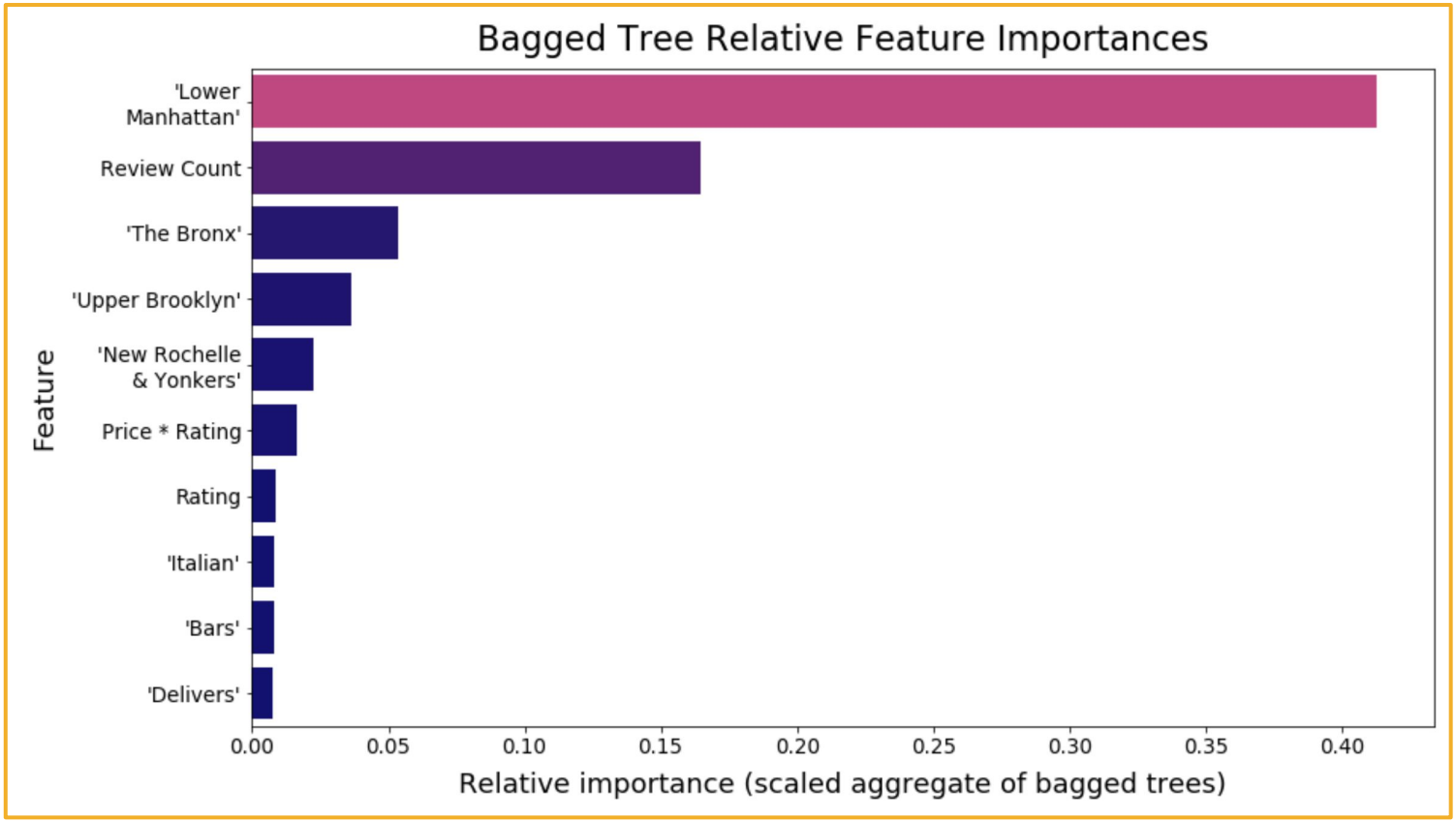
- **Target = Binary classification using 15% affluency threshold**

- **Model Scoring**
  - Not using accuracy
  - Reducing False Positives = Specificity

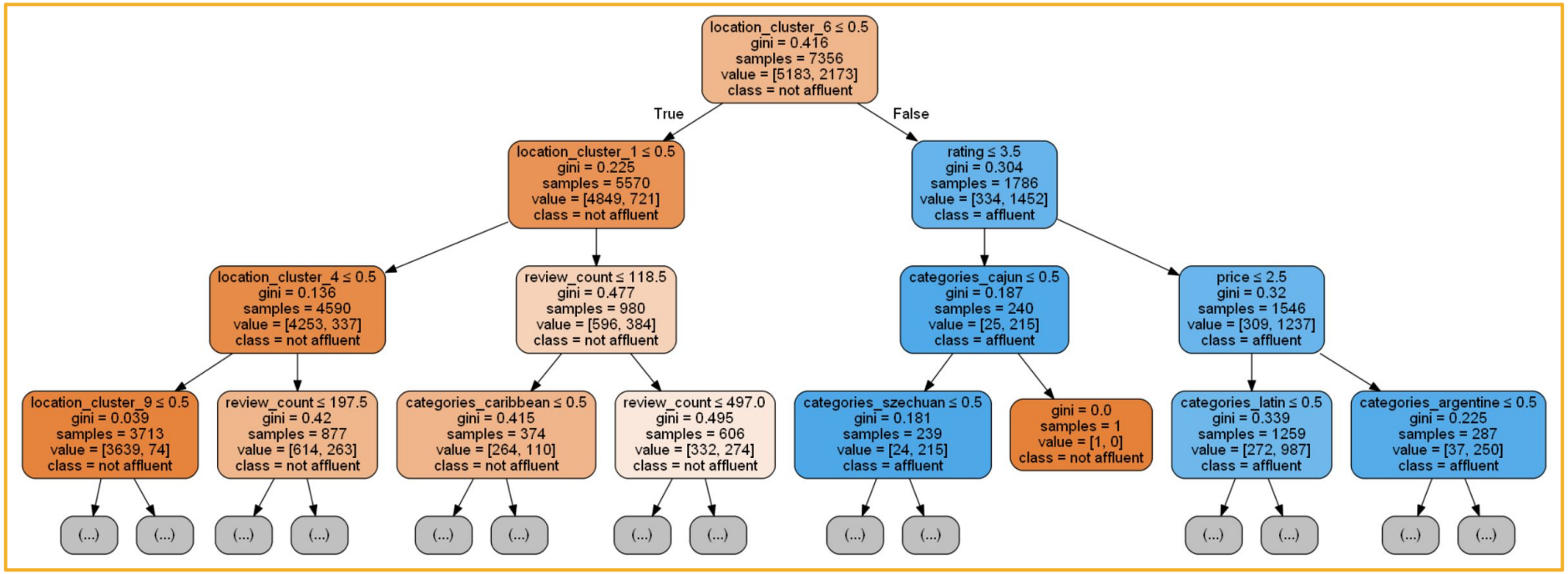
Model Selection

Model	Train Score	Test Score	False Pos. (%)	Accuracy
LogisticRegression	91.51%	90.34%	6.81%	85.45%
KNeighborsClassifier	90.47%	87.79%	8.60%	77.50%
DecisionTreeClassifier	93.58%	93.06%	4.89%	85.81%
BaggingClassifier	**99.46%**	**96.76%**	**2.28%**	77.33%
XGBClassifier #1	93.56%	93.06%	4.89%	85.85%
XGBClassifier #2	92.38%	91.61%	5.91%	85.69%

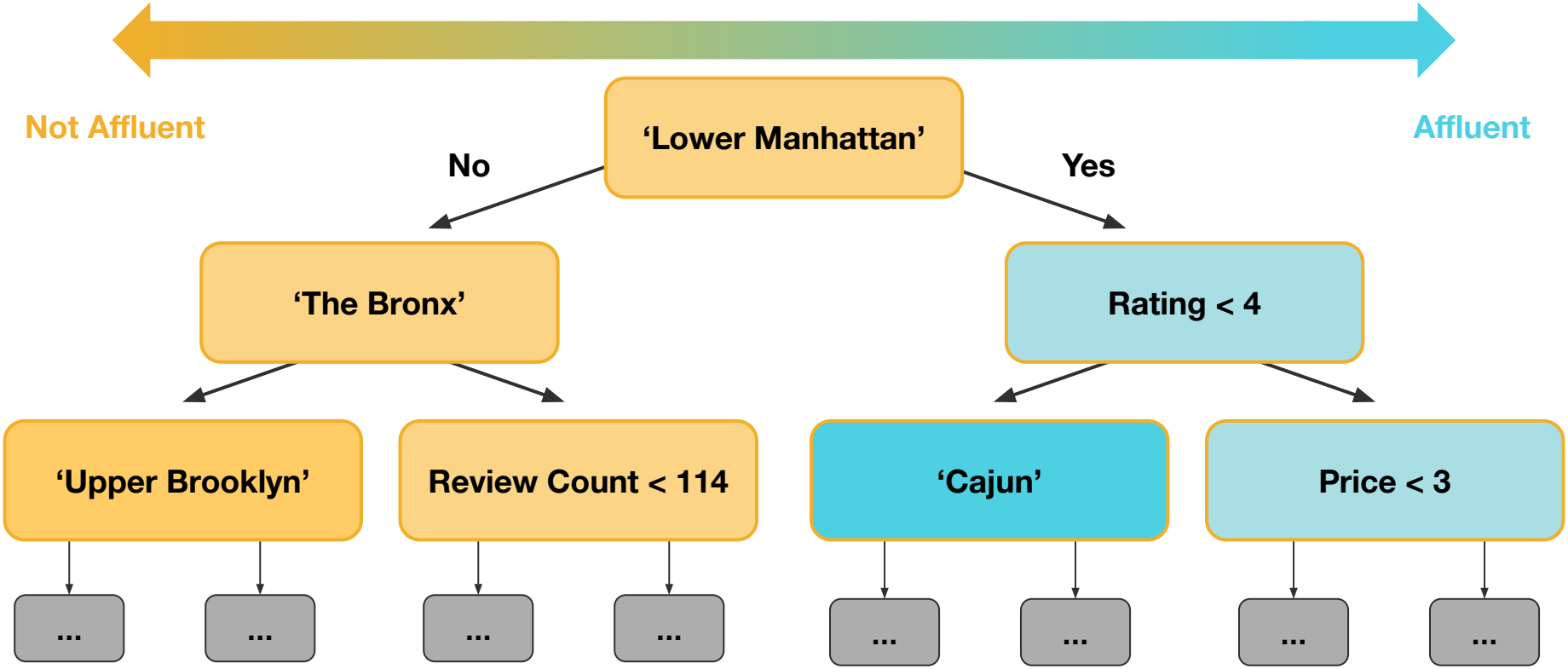
Model Evaluation



Visualizing Decision Tree



Simplified Decision Tree



Query Tool

Rockefeller Center

- **Making the model practical**
  - **Tool to quickly determine affluency for a particular area**

Known zip

	zip_code	affluency_rate	is_affluent
0	10001	0.161157	True

“Unknown” zip

	zip_code	affluency_rate	is_affluent
0	10020	0.277778	True

Conclusion

- **Yelp “$” Price alone was insignificant in determining affluency**
  - Not in our model’s Top 10 Important Features

- **Limitations**
  - Small area zip codes = few, if any, businesses
  - Removed businesses missing Yelp “$” price
  - Only works for NYC (sort of)

- **Assumptions of the model**
  - Affluence rate of 15%
  - Set $200k+ threshold
  - We assumed specificity was best

Recommendations

- Do not use Yelp “$” price to determine where to send emergency resources!
  - Location or # of Yelp reviews are much better predictors for affluence
- Yelp could include whether business is in commercial area or not

Future work:

- Population density as another focus, besides affluency
- Using # of reviews as the focus feature
- We would like to try making a model for other areas to test portability of our methodology

Source Documentation

- Yelp API - Business Endpoints
- IRS dataset
- NYC Zip Codes
- Super Severe Source