The
Erdős
Institute

# Foursquare Location Matching

Team Hopf Bundle:
Halley Fritze, Jay Hathaway, Max Vargas

# Motivation and Problem Statement

## Motivation:

- ❖ Businesses require reliable location information to run locals ads or expand to new cities.

## Problem:

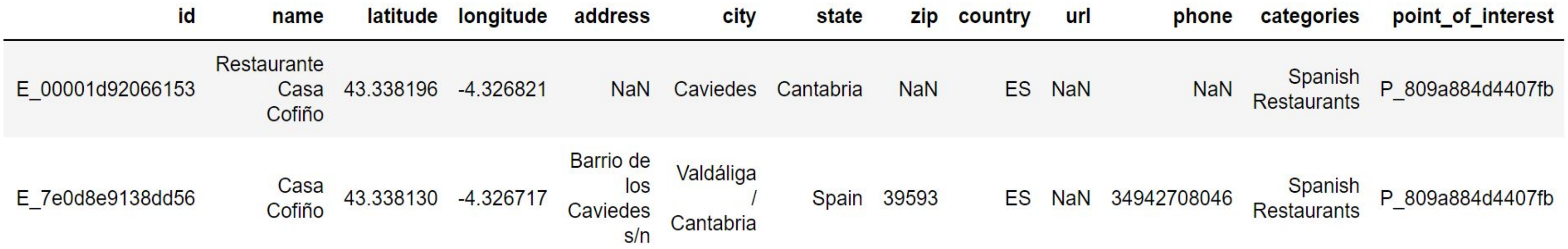
- ❖ These location data sets contain a lot of noise, unstructured information, and incomplete or inaccurate attributes.

## Goal:

- ❖ Match data points describing the same POIs using machine learning.

# Data Extraction

Foursquare is a location technology company which supplied our data via Kaggle.

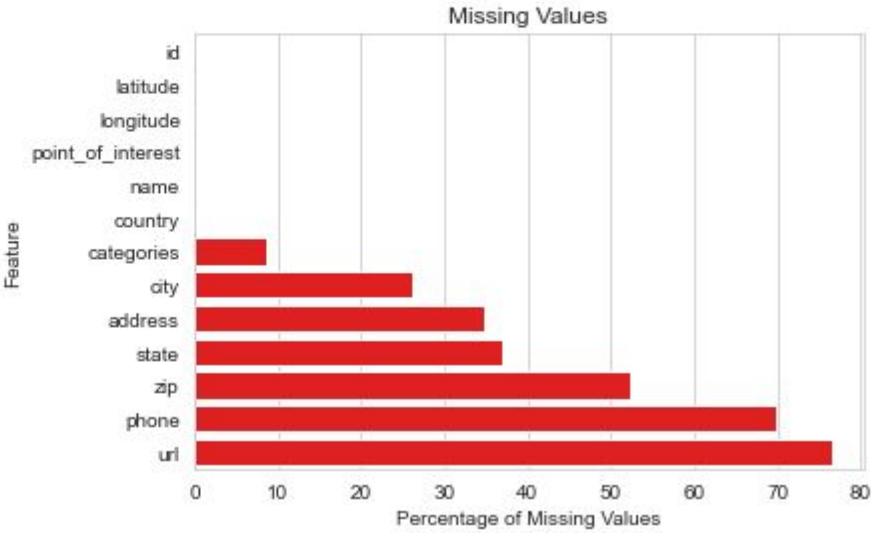


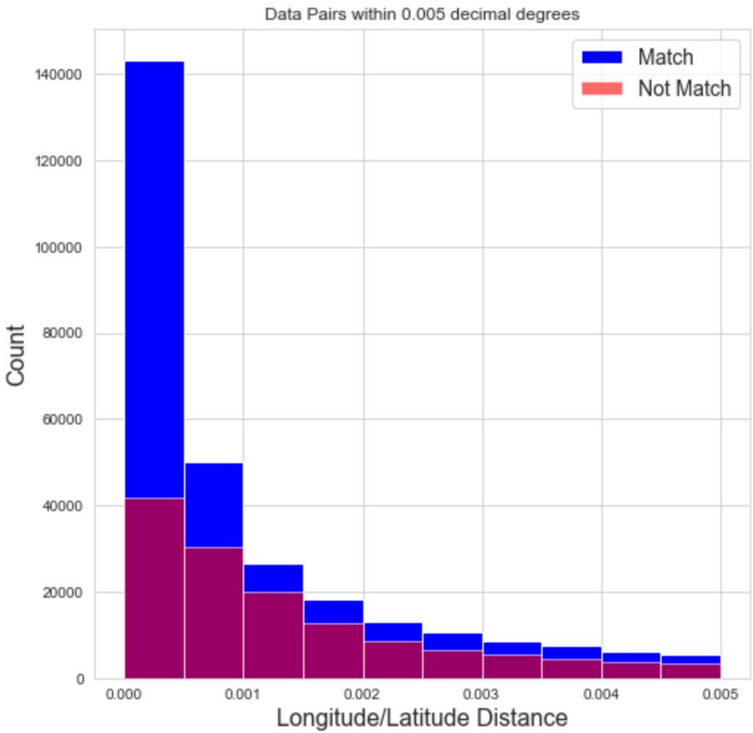
| id | name | latitude | longitude | address | city | state | zip | country | url | phone | categories | point_of_interest |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| E_00001d92066153 | Restaurante Casa Cofiño | 43.338196 | -4.326821 | NaN | Caviedes | Cantabria | NaN | ES | NaN | NaN | Spanish Restaurants | P_809a884d4407fb |
| E_7e0d8e9138dd56 | Casa Cofiño | 43.338130 | -4.326717 | Barrio de los Caviedes s/n | Valdáliga / Cantabria | Spain | 39593 | ES | NaN | 34942708046 | Spanish Restaurants | P_809a884d4407fb |

The above two data points represent the same POI.

Additionally, they supplied a data set containing pairs of points with a boolean ‘match’ feature.

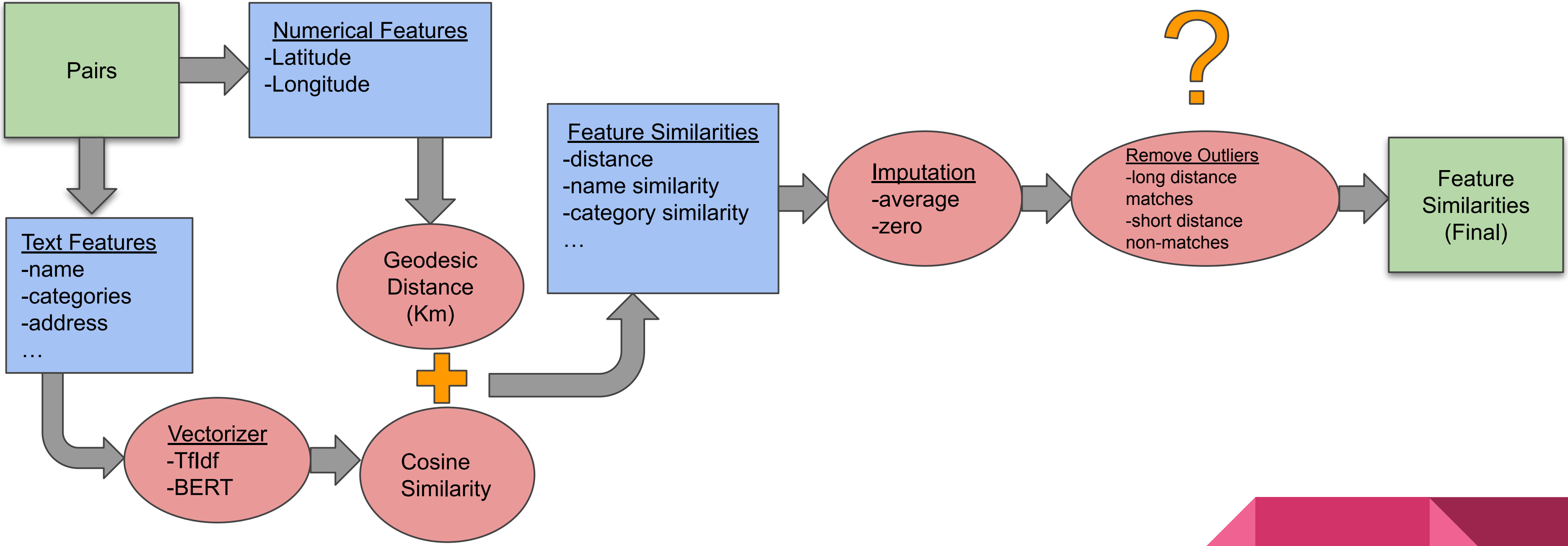
# Exploratory Data Analysis





- Many of our features are missing entries.
- Close distant pairs may not be the same POI.

# Feature Engineering



Pairs

Numerical Features
-Latitude
-Longitude

Text Features
-name
-categories
-address
…

Geodesic Distance (Km)

Vectorizer
-TfIdf
-BERT

Cosine Similarity

Feature Similarities
-distance
-name similarity
-category similarity
…

Imputation
-average
-zero

Remove Outliers
-long distance matches
-short distance non-matches

Feature Similarities (Final)

# Baseline Training

| | Accuracy | Precision | Recall |
|---|---|---|---|
| Logistic Regression (Distance) | 0.6889 | 1.0 | 0.6889 |
| Logistic Regression (Category+Name) | 0.7205 | 0.8994 | 0.7467 |
| K Nearest Neighbors | 0.7269 | 0.7612 | 0.8793 |
| Feed-Forward NN | 0.7259 | 0.9308 | 0.7390 |
| Random Forest | 0.7285 | 0.8644 | 0.7697 |

- ❖ Baseline models were trained on location, name, and category features.
- ❖ KNN, Neural Networks, and Random Forests achieved the best performances.
- ❖ Improvements depended on better data-cleaning techniques.

# Training with all features

| | Accuracy | Precision | Recall |
|---|---|---|---|
| K Nearest Neighbors | 0.7731 | 0.8010 | 0.8922 |
| Feedforward NN | 0.7771 | 0.9166 | 0.7924 |
| XGBoost | 0.7842 | 0.8998 | 0.8086 |

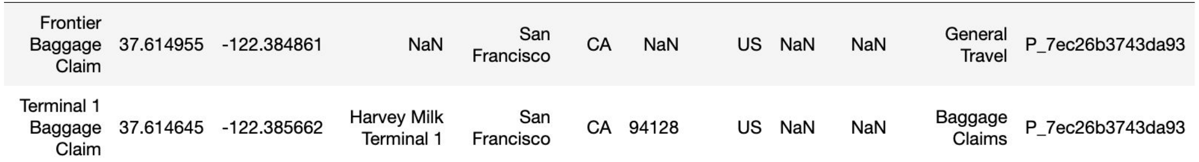
- Models were trained with all features in the dataset.
- XGBoost yields the highest accuracy.
- Missing data was imputed using mean values (+2% accuracy for XGB)

- “1 Towne Centre Blvd #2800” and “1 Towne Centre Blvd” have a TfIdf similarity of 0.8036.
- “400 Fairview Ave” and <NaN> have a similarity score of 0.5375 due to mean imputation.

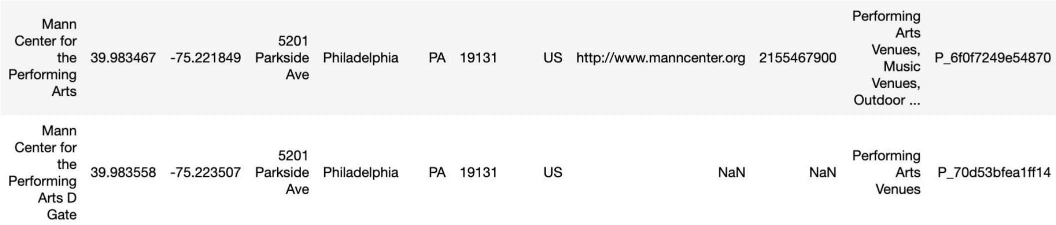
# Further Improvements

- Consider using BERT as our vectorizer for feature engineering.
- Develop better cleaning techniques to improve accuracy
- Analyze reasons behind false positives and false negatives
- Better zip-code processing

False Negative >



| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Frontier Baggage Claim | 37.614955 | -122.384861 | NaN | San Francisco | CA | NaN | US | NaN | NaN | General Travel | P_7ec26b3743da93 |
| Terminal 1 Baggage Claim | 37.614645 | -122.385662 | Harvey Milk Terminal 1 | San Francisco | CA | 94128 | US | NaN | NaN | Baggage Claims | P_7ec26b3743da93 |



| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Mann Center for the Performing Arts | 39.983467 | -75.221849 | 5201 Parkside Ave | Philadelphia | PA | 19131 | US | http://www.manncenter.org | 2155467900 | Performing Arts Venues, Music Venues, Outdoor ... | P_6f0f7249e54870 |
| Mann Center for the Performing Arts D Gate | 39.983558 | -75.223507 | 5201 Parkside Ave | Philadelphia | PA | 19131 | US | NaN | NaN | Performing Arts Venues | P_70d53bfea1ff14 |

< False Positive

Thank you to Akul and the Erdos Institute!