# Will a business close?

Yuxiang Gong


August 20, 2019

# Overview


## Project Description

## Feature Engineering

- Review and Check-in Information
- Enrich Features - population, land area, household income
- Location Clustering

## Machine Learning Model

- Sampling
- Model Selection
- Model Training

## Results and Insights

- Performance
- Feature Importance
- Further Improvements

## Summary

# Project Description

In the business dataset, about 18% of businesses have closed.

| | open | closed |
|---|---|---|
| Counts | 158525 | 34084 |

This project is going to get insights into Yelp's datasets to find the answers to the following questions:

- What could be the factors lead to the closure of a business?
- Can we predict if a business will stays open or closes in the near future?

## Idea:

Take the 'is_open' (1, 0) feature in the business dateset as the label, collect features to build a binary classification model.

# Review Information

The review data contains unique Id of businesses and users, text comments, comments entry time, stars and votes from other users. This project focuses on numerical data, thus the following features to each business are extracted:

- Total number of 'cool', 'funny' and 'useful' votes
- Number of reviews per month
- Duration of reviews in years (i.e. time interval between first time and latest comments)
- Year and month of first and latest comments

# Check-in Information

There are certain discounts to the users if they check-in a business on Yelp, which could be a detector of how much a business devotes to attract customers. The check-in dataset contains business Id and time points of check-in. Time related features are extracted:

- Total number of check-ins
- Check-ins per month
- Duration of check-ins in years
- Year and month of first and latest check-ins

# Enrich Features

Normally, a business is closely related to the residents around. However, Yelp can't provide these information. Thanks to **uszipcode**, a database which contains various features like geography, demographics, income in USA which can be searched via postal code, address and geographic coordinates.

In this project, the postal code is used to acquire statistics about residents around businesses in US, following information are obtained:

- Population and population density
- Land area
- Median household income

The number of businesses is reduced after searching through postal code in the US (from 192609 to 96632).

# Location Clustering

Apart from the local residents, the density of businesses might also play a crucial role in running a business. The latitude and longitude coordinates can be used to determine if a business is in a chain ($>5$) or not.

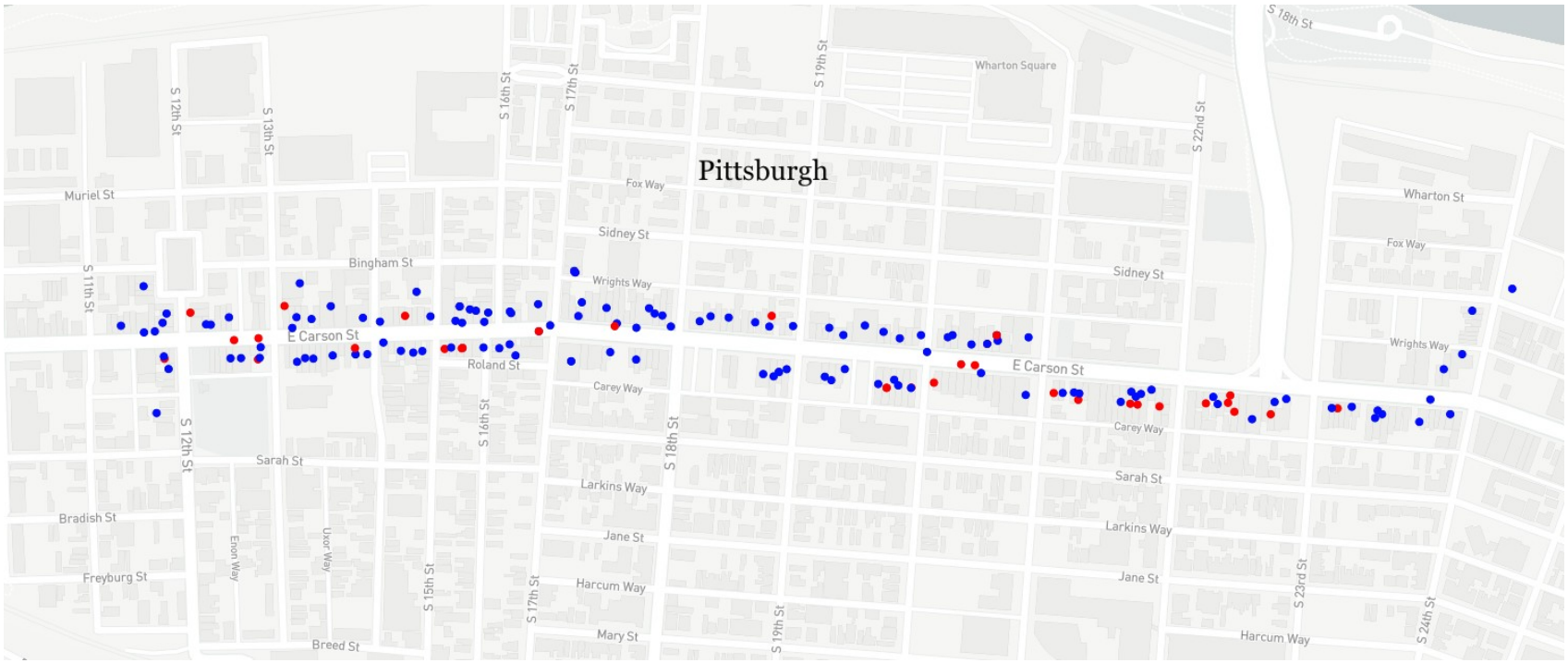
The DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm is used to cluster the businesses according to their latitude ($\varphi$) and longitude ($\lambda$). The distance are calculated with haversine formula:

$$hav(\Theta) = hav(\varphi_2 - \varphi_1) + cos(\varphi_1)cos(\varphi_2)hav(\lambda_2 - \lambda_1),$$

where $\Theta = \frac{d}{r}$, d- spherical distance, r- radius of the earth.

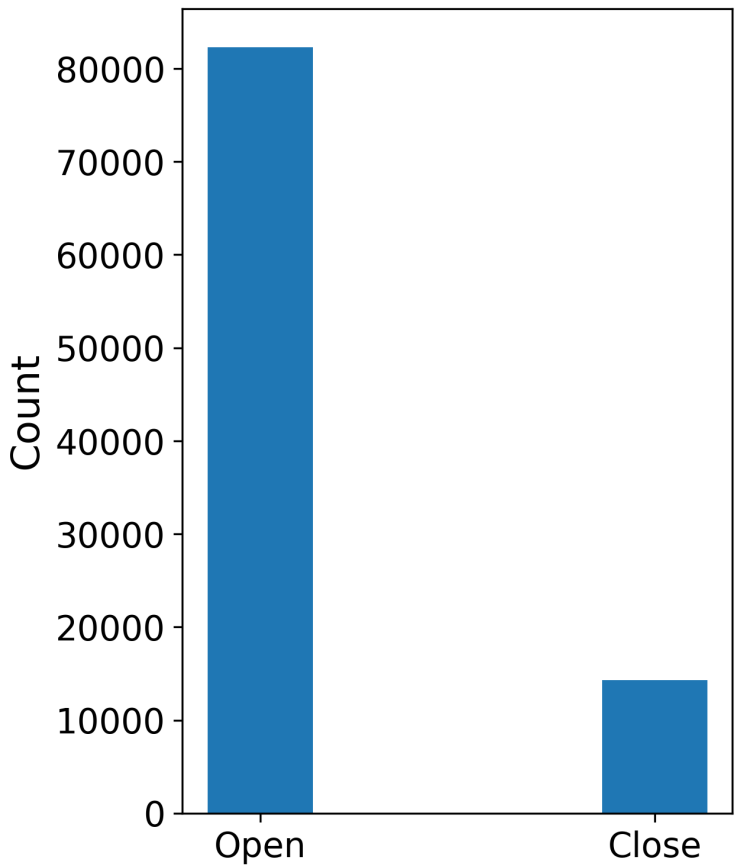
# Location Clustering

If the minimum distance between two businesses is within 50 meters, the businesses are clustered in a group. Here is the result of a cluster in Pittsburgh, PA. There are 171 businesses (35 closed and 136 open) on the same street, which are supposed in a chain.



# Sampling

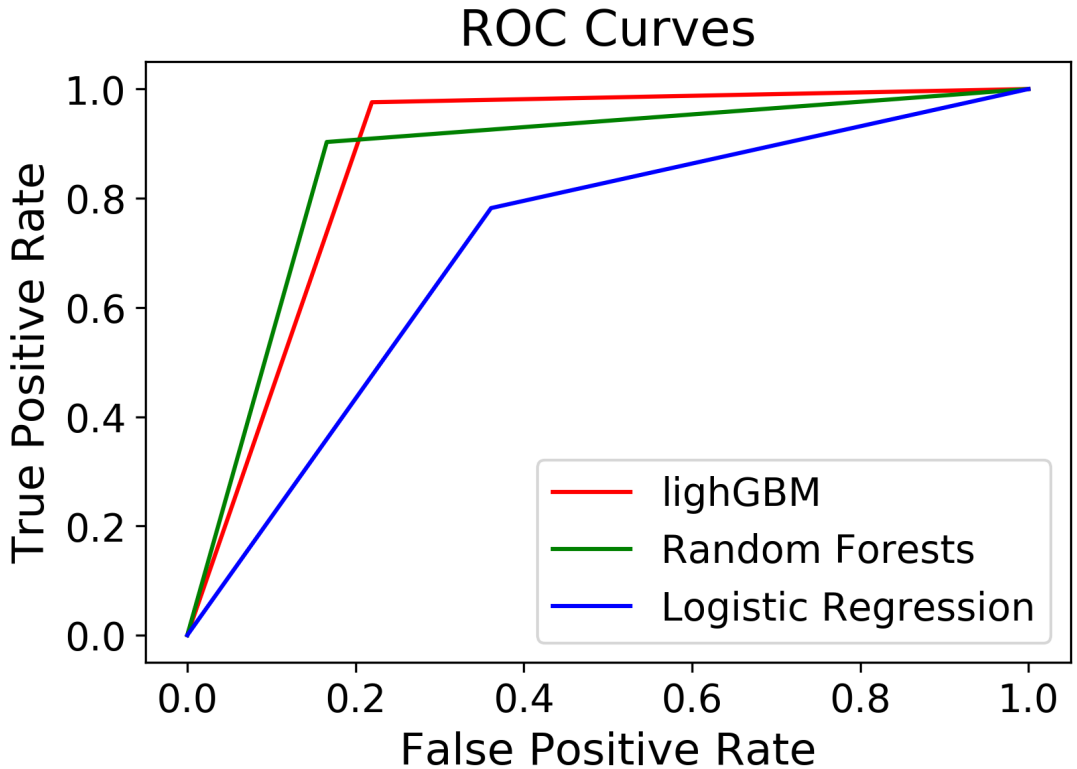
The number of closed businesses is about 15% after feature engineering, which would cause over-fitting on the open businesses due to the imbalanced datasets. The solution is to oversampling the closed businesses to comparable amount of the open businesses.




- Split training and testing data (33%)
- Randomly oversampling the closed businesses in the training set to 80% of the open businesses

# Model Selection

The features like population, year, cluster labels are in different scales, thus a model which is not sensitive to the scale is preferred. Among the commonly used algorithms for binary classification, lightGBM, Random Forests and Logistic Regression are employed. Here is the performance in ROC curves. **lightGBM** has the best performance.



# Model Training

The sampled training data is passed to lightGBM, and the prediction is made with the testing data.

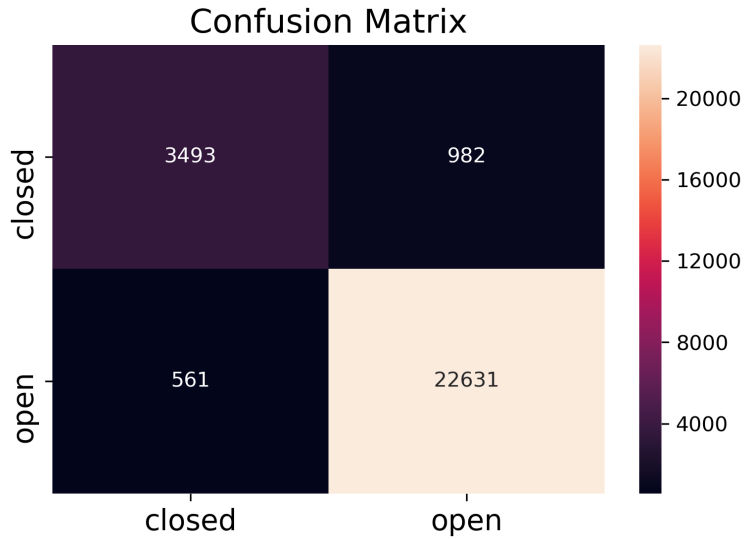
The hyperparameters are tuned with RandomizedSearchCV, with the scoring metric as recall. The function can be activated by assigning tune_parameters=True while running the code.

```
# run model training
model.main(df_business, tune_parameter=False)
```

# Performance

Overall accuracy: 94.42%

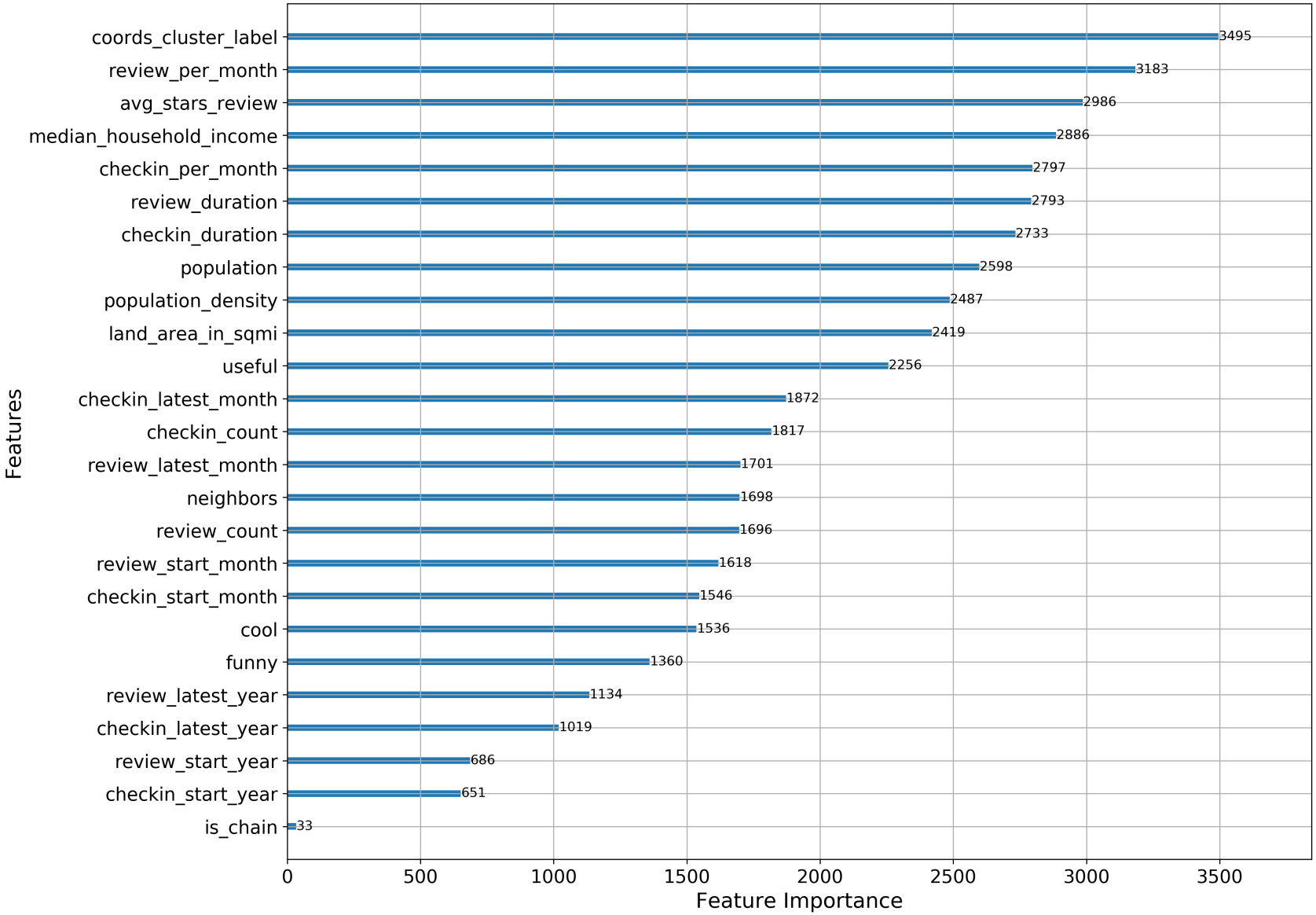
| | open | closed |
|---|---|---|
| Precision | 0.96 | 0.87 |
| Recall | 0.98 | 0.77 |
| f1-score | 0.97 | 0.82 |
| Support | 23192 | 4475 |




As demonstrated above, the precision of open business is 96% and closed business is 87%. This means that among the businesses that are recognized as open by the model, 96% of them actually remained open. The remaining 4% are false positives. As to the closed businesses, 87% are correctly recognized.

For a business investor, the interested factor could be the recall which includes the factor of how many businesses are actually closed but are not predicted as closed.

# Feature Importance



# Feature Importance

According to the feature importance that resulted from this model, following insights can be concluded:

- The most important feature is the location clustering of the businesses
- Review rate, duration and stars also play essential roles
- Check-in factors which indicate how much effort a business is put on attracting customers are important as well
- The local residents' situation including income, population and density are crucial

These factors fit well with our common sense about if a business will success or not. In conclusion, the model performs well in predicting the status of a business.

# Further Improvements

Due to the imbalanced datasets, the training procedure tend to overfit the data after oversampling. The failure of a business is complicated, additional features could be taken into consideration to further improve the performance of the model:

- Rental or housing price of the local area
- Price of the services
- Distinguish the type of a business (i.e. restaurant, SPA, bar, etc.)
- Analyze the review texts (NLP)
- Collect more data related to closed businesses

# Summary

- Classify businesses in the US with Yelp's data into open and closed with lightGBM
- Information of local residents are introduced according to zipcode
- Businesses are clustered with latitude and longitude
- An overall accuracy of 94.42% is achieved
- Due to the imbalance of open and closed datasets, the model has high variance on the closed businesses