Predicting Business Closure With Supervised Learning

Objective

Predict business closure using Yelp data. This type of information can be useful to owners as an early warning that their business might be at risk.

Data Source

The Yelp Dataset, a publically available subset of their business, review and user data. Downloadable as a SQL archive, it can be loaded directly into a MySQL database in the AWS cloud.

Data Selection

The data set covers many cities and businesses but for the purpose of this project it was narrowed down to:

- Las Vegas, the city with the most businesses and reviews
- Restaurants, the business category most likely to be reviewed
- the most recent 3 years of individual reviews

The end result is a set with 4085 businesses, out of which 795 are closed.

Features

Feature	Description	Indicator	Engineered	Keep
bpnr_cnt	Number of similar businesses/neighborhood	🚫	✅	✅
cat_...	Category	✅	🚫	🚫
nbr_...	Neighborhood	✅	🚫	🚫
rating	Global -1/0/1 based on stars binning	🚫	✅	✅
review_count	Total number of reviews	🚫	🚫	✅
review_sentiment	Most recent N/% reviews, converted to -1/0/1 sentiment and summed up	🚫	✅	✅
same_name_cnt	Number of businesses with the same name (chain?)	🚫	✅	✅
stars	Global number of stars	🚫	🚫	🚫
zip_	Postal code	✅	🚫	🚫

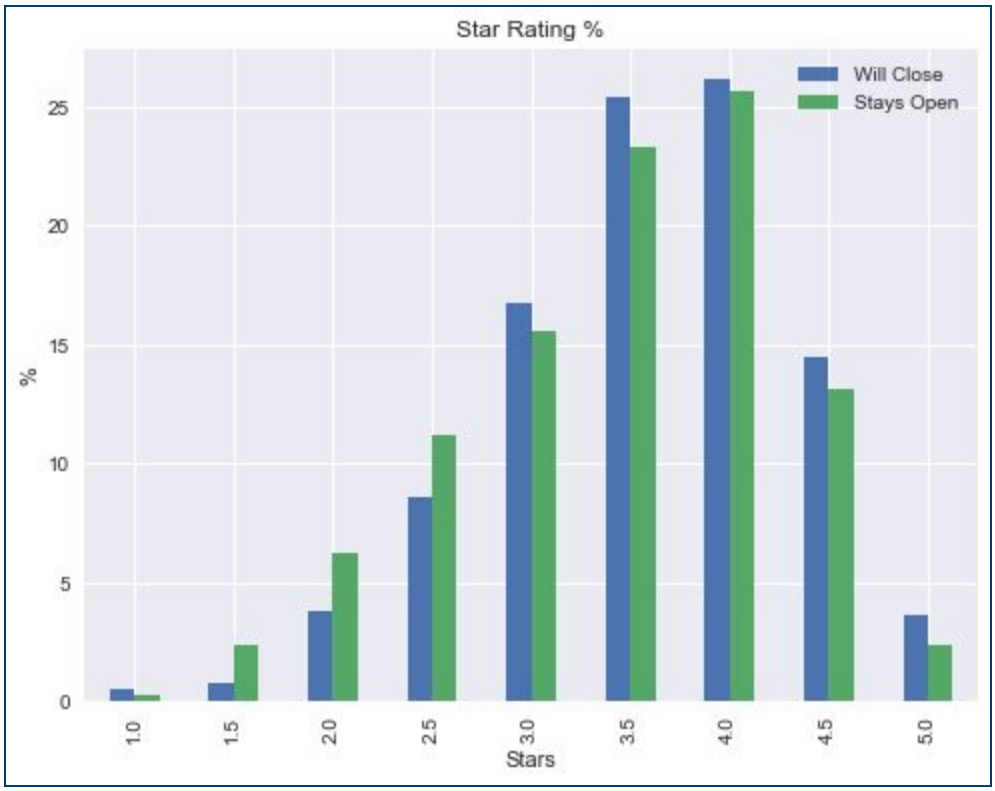
Target And Score Selection

Target: **will_close**, binary choice
Scores:

- **Recall**: the most important score, false positives (business flagged yet not closing) are less risky than false negatives (business not flagged, yet closing)
- Precision
- F1
- AUC

EDA On Stars Rating Relevance

Stars rating does not have much prediction power since the distribution is very similar between closed and open businesses.



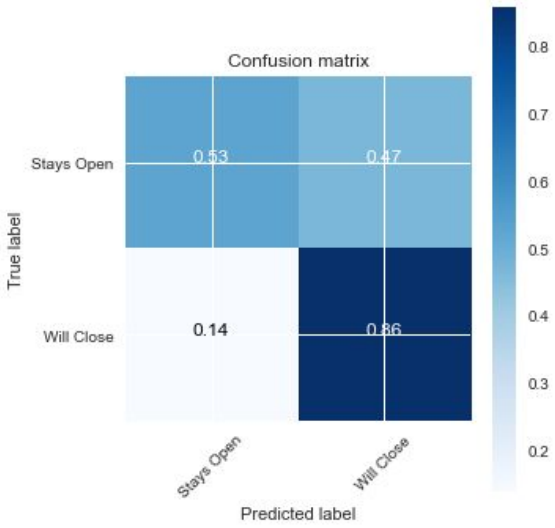
Feature Trials

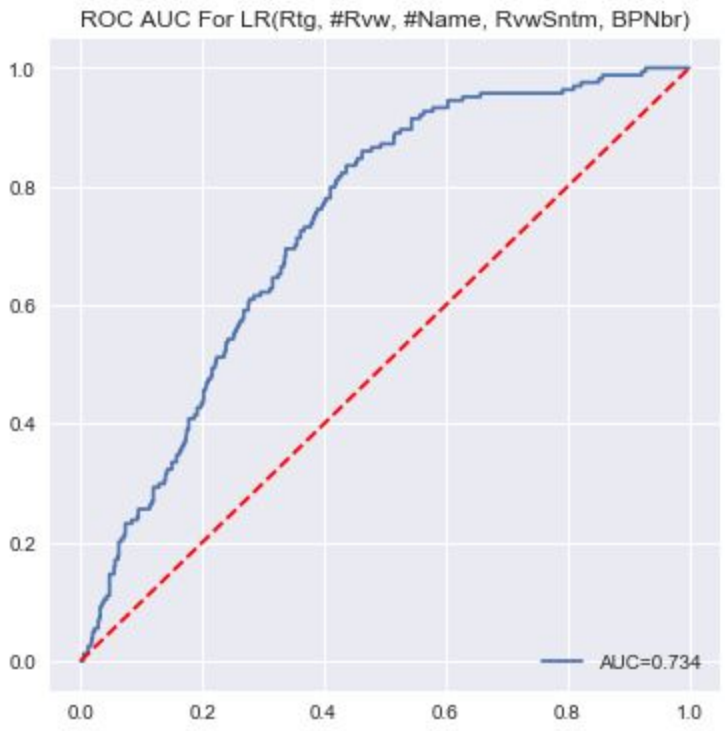
- The data set was split 80/20 into train/test subsets.
- The 2 target classes are imbalanced ⅛ (positive/negative); auto class weighting was used for all models.
- Logistic Regression was run with default regularization C=1.

- Different sets of features and models were tried and the following observations were made:
    - Location (ZIP, neighborhood) and categories do not help
    - Logistic Regression has better recall than RandomForest
    - Precision is low

Final Results On The Test Set

Recall	0.860
Precision	0.314
F1	0.460
AUC	0.734





Conclusions

- The model could be deployed as a Web App to be used by owners to check if their business is at risk of closing
- Unfortunately the current implementation has very low precision, “cry wolf” syndrome would lead to dismissal
- Possible refinements:
    - NLP for review and tips parsing
    - Better feature engineering
    - Include economic data, e.g. changes in lease or energy costs
    - Include demographic data