A Machine Learning based Yelp Recommendation System

Pranav Bhardwaj[1], Nicolas Bievre[2], Frederik J. Mellbye[3]

pranavb@stanford.edu, nbievre@stanford.edu, frederme@stanford.edu

Department of Statistics, Stanford University[1,2]

Institute for Computational and Mathematical Engineering, Stanford University[3]

Abstract

The Yelp Open Challenge data set contains user, business, and review data from 2004 to 2018. Binary classification is performed using user and business features to predict positive labels of 5 star ratings. High predicted probabilities correspond to recommendations. Linear and tree based models were used. Tree based models performed well due in part to their ability to represent nonlinear relationships. Models generalized well to the future.

Data

Each row corresponds to a review containing user and business information. The label y is positive (1) if the user u left the business b as 5 star review, and negative (0) otherwise.

Table 1: Train-Validation-Test Split

	Reviews	Users	Businesses	5 ★
Train	4,903,362	1,259,558	125,159	43.1%
Validation	527,276	262,764	75,605	51.5%
Test	527,276	274,722	75,254	51.8%

Features

Input information to learning algorithms, x, is a concatenation of the user features, x_u, and the business features x_b. A single example i is then:

$$\begin{bmatrix} x_u^{(i)} \\ x_b^{(i)} \end{bmatrix} = x^{(i)} \in \mathcal{R}^d, y^{(i)} \in \{0,1\} \quad (1)$$

Where $d = 166$ is the number of predictors.

Models

Logistic regression:

In logistic regression, the predicted probability of a 5 star rating is of the form:

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1+e^{-\theta^T x}} \quad (2)$$

Models are trained using L_1 and L_2 regularization with unit regularization strength.

Gaussian Discriminant Analysis (GDA):

In GDA, it is assumed that $p(x \mid y)$ follows a multivariate normal distribution. Model parameters ϕ, μ_0, μ_1 and Σ are fit by maximizing the log-likelihood of the data given by:

$$\ell(\phi, \mu_0, \mu_1, \Sigma) = \log \prod_{i=1}^{n} p(x^{(i)} \mid y^{(i)}; \mu_0, \mu_1, \Sigma) p(y^{(i)}; \phi) \quad (3)$$

A linear decision boundary at which $p(y = 1 \mid x) = 0.5$ is created.

Tree based methods [1]:

Decision Trees: In Decision Trees, the input space $\mathcal{X}$ is repeatedly split into two child regions by thresholding a single feature. The work presented uses cross-entropy loss, which is of the form:

$$L_{cross}(R) = -\sum_c \hat{p}_c \log_2 \hat{p}_c \quad (4)$$

The predictor split corresponding to the maximum reduction in loss is made at each step.

Random Forests: Random Forest Classifiers are constructed by bagging Decision Tree Classifiers which are trained using a random subset of features. To predict on a new example, the majority vote from the Decision Tree Classifiers is returned.

Adaboost: AdaBoost classifiers sequentially apply base classifiers $G_m(x)$ (here decision trees) to modified versions of the data. The initial version of the data has uniform weighting. In each successive iteration m, observation weights are modified to place more weight on examples misclassified by the previous classifier $G_{m-1}(x)$.

Results

Model selection was based on validation set accuracy, reported in Table 2.

Table 2: Train and validation accuracy, implemented in *scikit-learn* [2]

	Train	Validation
L_2-regularized Logistic Regression	63.73 %	60.53 %
L_1-regularized Logistic Regression	74.15 %	74.84 %
GDA	74.07 %	74.69 %
Decision tree	75.37 %	75.97 %
Random Forest	71.58 %	71.56 %
AdaBoost	75.33 %	76.13 %

The final model selected was an AdaBoost classifier with 40 base estimators. Each base estimator was a decision tree classifiers of maximum depth 4. Test set performance is summarized in Figure 1 and Table 3.

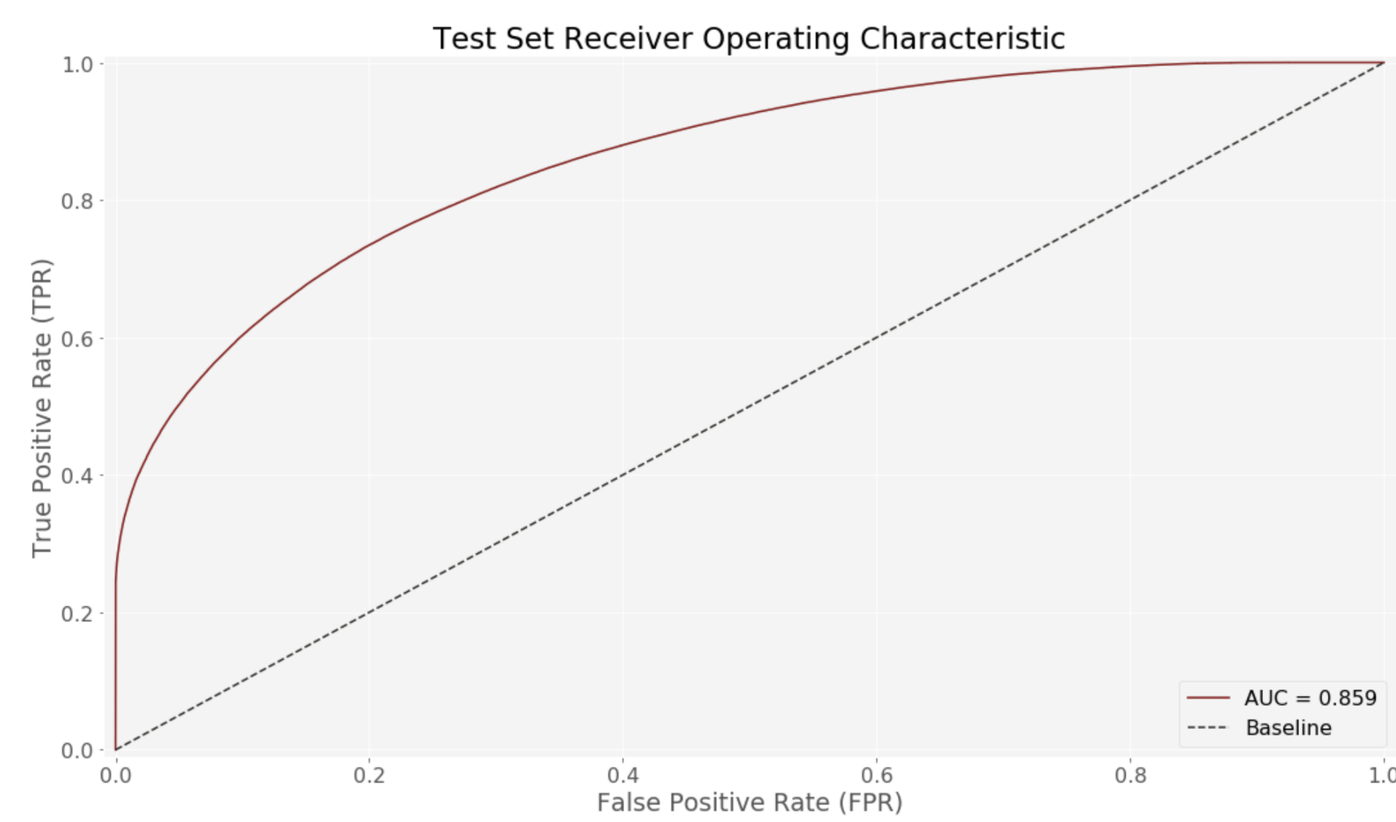

Figure 1: Test set AUC-ROC curve for the AdaBoost classifier

Table 3: AdaBoost classifier test set performance

	Accuracy	Precision	Recall	F1
AdaBoost	76.63 %	79.18 %	74.53 %	76.78 %

Discussion

Logistic regression with L_1 regularization as well as tree based methods performed and generalized to future data well. We believe this is because these models effectively use a subset of predictors, and many predictors are not adding useful signal. This would also explain the poor performance of logistic regression with L_2 regularization. After cross-validation for model selection, an AdaBoost classifier with 40 base estimators, each a decision tree classifier with maximum depth 4, was deployed on the most recent data available. This model performed even better on test data than validation data.

Future Directions

- Perform more extensive feature engineering
- Leverage more of the available Yelp data (tips, review text, photos)
- Leverage the community structure in Yelp with collaborative filtering and graph based models
- Provide recommendations to businesses on improvements they can make

References

[1] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference and prediction.* Springer, 2 edition, 2009.

[2] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.