Yelp Rating Prediction Visualization

Jiawen Liu
Keke Wu
Wei Miao
Xu Han
Yawen Zhang

ABSTRACT

Yelp is a crowd-sourced local business review and social networking community, which has hundreds of thousands of users contribute their data every day. Based on users' reviews and ratings, good local businesses stand out among their categories on top of the list, acting as a word-of-mouth reference. However, tons of user data doesn't make any sense unless we make good use of it. In this case, we decided to look into the problem that how we might get as much as useful information from our data with a particular interest in how the users' rating behavior are influenced by different factors, and what kind of prediction we can make out of the rating pattern we found. With various features extracted from Yelp data, we conducted feature selection, best split and finally built a Decision Tree model which predicts users' rating based on those features. Our visualization includes the complete process of this typical machine learning method, which provides insights about the inner mechanism of how the rating prediction is conducted.

KEYWORDS

Yelp; Rating Prediction; Decision Tree

ACM Reference Format:
Jiawen Liu, Keke Wu, Wei Miao, Xu Han, and Yawen Zhang. 2018. Yelp Rating Prediction Visualization. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Rating prediction plays an important role in the recommendation system as to capture users' preference for specific product. In Yelp, user can post their rating and review for a restaurant they visited, and these data as well as user/restaurant related attributes, e.g., user average rating, restaurant location or category, can be well gathered for conducting the rating prediction. If a user's rating for a restaurant can be accurately predicted, Yelp would be able to recommend high rating product to the user which meets their preferences.

The challenge in this problem lies in the feature selection and modeling processing. With multiple features related to a user's rating, picking up the effective features is quite important, which would significantly influence the prediction performance. Additionally, the modeling process which selects the best machine learning model for rating prediction is also critical, and it's necessary to gain insights about how the model functions, i.e., the inner mechanism of model. Previously, most machine learning processes have been conducted in a "black box", by just showing input and output. In our project, we aim at visualizing the whole process, including feature selection, best split, and model tree. The Decision Tree model is selected because its straightforward idea in classification, which would be beneficial for visualization.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 RELATED WORK

In this project, we deploy the visualization principles and techniques to make mechanism of the whole recommendation system(RS) transparent based on Yelp's public dataset[]. A lot of research has been done on recommendation system and the RS techniques are broadly divided into two types: memory-based approach, which recommend business based on similarity or correlation between users[], and model-based approach, which use machine learning methods to predict user ratings[]. In our project, we use decision tree from model-based approach[] as our visualization example.

Our visualization of modeling process mainly focus on four parts: feature engineering, best split analysis, feature ranking and model training. For feature engineering, we extract 22 features in total, includes user-related features(7), business-related features(3), user-category features(5) and review-related features(7). Among review-related features, we extract several advanced natural language processing(NLP) features like polarity[] and subjectivity[]. For best spllit analysis, we create a moveable threshold to study how this feature – business average star, could influence the decision making(whether to recommend or not). When moving the threshold, the calculated accuracy and true positive rate[] will be changed correspondingly and we can choose the threshold with highest accuracy as our decision tree's best split. In the part of feature ranking, we measure the importance of all 22 features based on the score retrieved by Xgboost[]. Xgboost's feature importance method calculateds F score, which indicates how many times the feature split on. Higher the F score is, more important the feature is. In our feature ranking visualization, we use a radar graph to show the importance of all these features based on F score. The last step is model training, we use 100 users as an example. 80 users are used to train and 20 users are used to test. The top three features with highest F score are selected and used by the model. We visualize each users path and the overall test accuracy.

3 DESIGN PROCESS

During the design process, we tried to get our results displayed in a reader-driven rather than author-driven way, following a story-telling style. Our preliminary works included data cleaning, data organization by categories, feature selection, hand-written sketches of the complete workflow. We used web as our platform and D3.js as the main tool. The web was divided into six different but consistent sections in the order of Our Story, Our Goal, Our Design, Visualization, Split and Tree Model, which was exactly the sequence we were following to dive into this topic. In terms of the visual design, we used the red, black and white color scheme adopted by Yelp to make it look consistent. Besides text descriptions and interactive visualization graphs, we also had hand-written sketches on the web as a proof of concept, which helped enrich the design elements' diversity. In terms of the visualization design, we used four different types of charts to visualize four different rating related data features. The restaurant ratings by category were visualized in a bar chart, showing customers' preferences on cuisines. The average ratings by state were visualized in a US map, with the darker colored state having a higher rating trend. We also extracted key words from users' reviews and got them displayed in a randomly generated word cloud, where the words were extinguished from each other by five colors mapped with five different rating stars. Besides these, we also built a network to visualize the influence on ratings from Yelp users who have most friends. All of these charts were interactive and dynamic, enabling a good user-directed exploration experience. Following sections utilized a split framework to explain how we applied machine learning to help us make predictions. Finally, a decision-tree like model walked the audience through the decision and prediction making process. With a well-designed and fully-interactive design style, we hope to help our audience get valuable information in a way that can be adjusted as they like. Meanwhile, we hope to use some vivid graphs and animations to bridge the gap between scientific research and general public understanding, making the abstract concept intuitive and straightforward.

4 MODELING PROCESS

For Yelp rating prediction, we use Decision Tree, which is tree-like graph of decisions and their possible consequences. It is widely used prediction model in machine learning. In this tree structure, leaves represent class labels and branches represent conjunctions of features that lead to the class labels. A typical example is shown in Figure 1 [1]. The modeling process include feature selection, best split, feature ranking and tree generation.

Feature Selection. Yelp dataset contains 11 tables such as Yelp dataset contains 11 tables such as business, category, checkin, etc. Figure 2[2] shows the dataset structure. We extract 22 features(shown in Table **??**, which could be divided into four groups: user-related features(7), business-related features(3), user-category features(5) and review-related features(7).

Best Split. Decision Tree works in a top-down manner, we need to choose a variable at each step that best split the sets of items.

[1]https://en.wikipedia.org/wiki/Decision_tree_learning
[2]https://www.yelp.com/dataset/documentation/sql

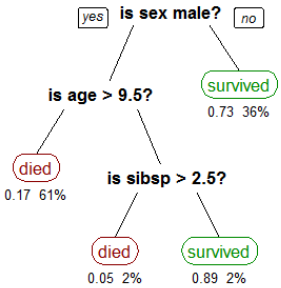

Figure 1: A Decision Tree showing survivals of passengers of the Titanic.

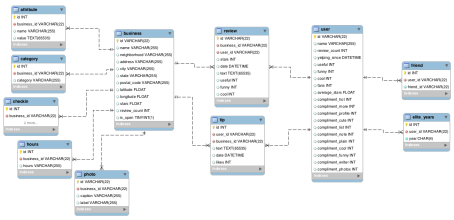

Figure 2: The structure of Yelp Dataset

Table 1: Feature Groups

Feature Group	Features
User-related Features(7)	user compliment count user funny count user cool count user useful count user fans count user review count user average rating
Business-related Features(3)	business average rating business location-related rating business neighborhood-related rating
Review-related Features(7)	Polarity Subjectivity TF-IDF Meaningful word count review useful count review cool count review funny count
User-category Features(5)	uc-review count uc-average rating uc-review funny count uc-review cool count uc-review useful count

The best split are chosen by certain metric, e.g., Gini impurity, Information gain and so on.

Feature Ranking. Feature ranking helps us have a better understanding of how each feature contribute to our model training. Table 2 shows the result of our 22 feature ranking. From this table we can know that review's polarity, user to a certain category's average rating, user to a certain category's review count.

Table 2: Feature Ranking Results

Feature	F score (scaled)
polarity	85
uc-average rating	67
uc-review count	42
subjectivity	42
business average rating	29
user average rating	26
user useful count	16
review useful count	16
TF-IDF	16
user review count	16
user compliment count	14
review cool count	11
review funny count	10
uc-review useful count	8
user funny count	8
user fans count	6
user cool count	4
uc-review cool count	2
uc-review funny count	1
business location-related rating	1
business neighborhood-related rating	1
meaningful word count	1

Tree Generation. With the selected features, and variables chosen by best split, a tree can be generated. And each path in the tree represents a decision-making process, in which a user's rating can be predicted based on a series of decision makings. Here we use only 80 users to train our model. Under this situation, even though the generated decision tree has really simple structure and only top three features are involved, the trained model could still have high prediction accuracy(100%).

5 RESULTS

Our visualization storyline include four parts, features, best split, feature ranking and tree model.

5.1 Vis 1: Features

In this part, as shown in Figure 3, we selected four typical features related to users' rating, including 1) restaurant type, visualized with a bar chart, 2) restaurant location, visualized with a map, 3) frequent words related to different ratings, visualized with a word cloud, 4) friendship network, visualized with a network. With these four selected features and different visualizations, we aimed to illustrate the difference of features. In each visualization, we added the rating information by either using the tooltip or other methods to show how the features are related to user's rating.

Figure 3: Visualizations of selected features related to users' rating.

5.2 Vis 2: Best Split

As to demonstrate how different cut-offs would influence the prediction results, we used user's average rating as an example, to illustrate the influence of thresholds on correct and other rates in the prediction results. The best split is determined by choosing the right threshold of the feature. As shown in Figure 4, when the threshold of user average rating is changed, the correct, incorrect, true positive, positive rate would all change. For all the features used for rating prediction, their best splits would be determined in advance.

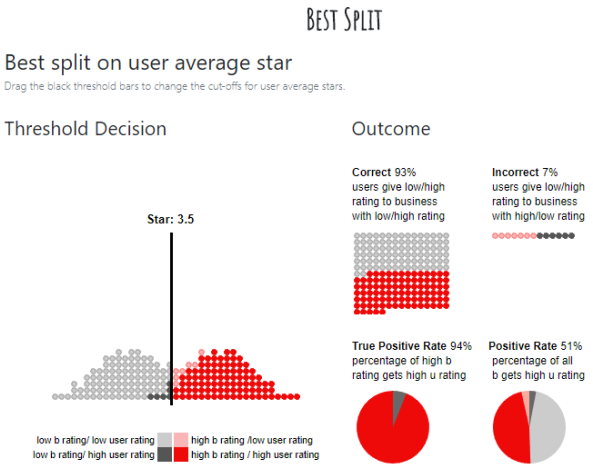

Figure 4: Visualizations of how to choose best split.

5.3 Vis 3: Feature Ranking

As shown in Figure 5, with the feature ranking results shown in Table 2, we further visualize the result using an radar map. This map illustrate the importance of different features in rating prediction. This is an important process before generating the tree model as only high ranked features would be used in the tree model.

5.4 Vis 4: Tree Model

Finally, with selected features from ranking, we built the decision tree model for rating prediction as shown in Figure 6. This visualization include the basic tree model with nodes represent the class labels and edges represent conjunctions of feature that lead to the class labels. This tree map works in a dynamic way of showing how a given input would be classified, i.e., how a user's rating would be

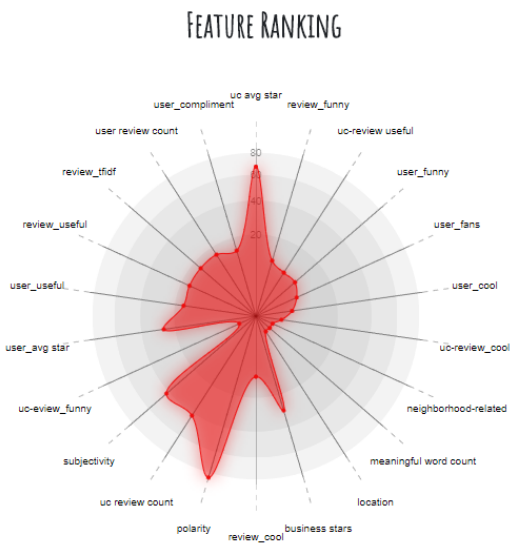

Figure 5: Visualizations of feature ranking result.

predicted given related features. Additionally, the training and test accuracy are also calculated.

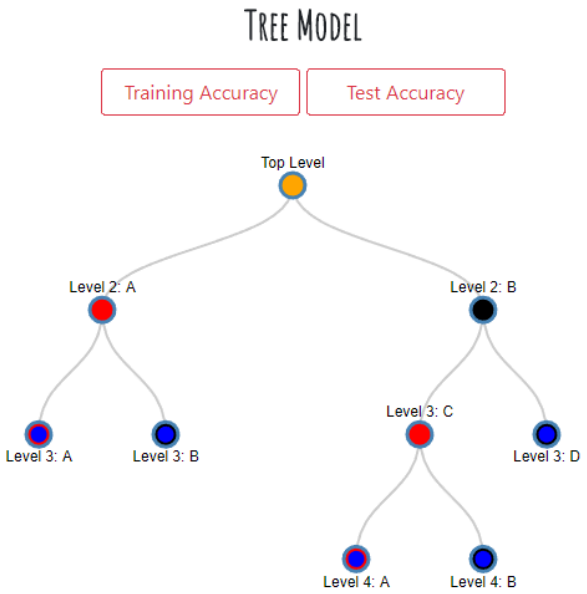

Figure 6: Visualizations of tree model and training/test accuracy.

6 DISCUSSION

In the project, we conducted a visualization of Yelp rating prediction. This visualization not only visualize the data but also visualize the machine learning process, as to reveal information in the "black box". With four parts of visualization, 1) features, 2) best split, 3) feature ranking, 4) tree model, we can understand the Decision Tree model better and gain insights of each part through visualizations.

REFERENCES