Yelp Rating Prediction Visualization

Jiawen Liu
Keke Wu
Wei Miao
Xu Han
Yawen Zhang

ABSTRACT

Your motivating problem, what you did, and what you found

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, **??** pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The motivating Problem, why it's interesting or important

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.

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

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.

2 RELATED WORK

Summarize research related to your projects, minimum 8 citations.

3 DESIGN PROCESS

Justifications for any design elements

During the design process, we tried to get our results displayed in a reader-driven rather than author-driven way, following a storytelling 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

Decision Tree

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]. The modeling process include feature selection, best split and tree generation.

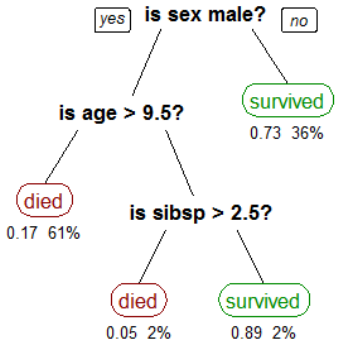

Figure 1: A Decision Tree showing survivals of passengers of the Titanic.

Feature Selection. There are multiple features related a user's rating. give list or table, showing all features.

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

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.

5 RESULTS

Our storyline

Our visualization storyline include three parts, features, best splits and decision tree model.

[1] https://en.wikipedia.org/wiki/Decision_tree_learning