Capstone Project Proposal

Hotel Feature Labeling and Rating Prediction

AI-JHE (Tom), LIN - 01/31/2018

Domain Background

Text topic modeling and sentiment analysis have long been interesting tasks for many researchers, for instance, the widely mentioned task on fake news detection or the case of product reviews. In this project, I'd like to leverage the Hotel Review dataset collected from Booking.com to answer two questions. One is of all the traits on Hotels, what trait best describes the impression customer has on every hotel. The other is the experiment on predicting customers' hotel rating based on their review text and other historical information.

When dealing with hotel reviews, most analysts implement techniques such as EDA, Geospatial Visualization and Bag of Words Modeling[1]. In my project, I will follow up the steps mostly, but elaborate more with some statistical techniques. One is the **Latent Dirichlet Allocation**, which is for the hotel feature labeling and the other is **the rating prediction model incorporating of external features beside review text (further discussed in Benchmark Model)**, hoping it would improve the precision for attitude rating prediction.

Problem Statement

In the world of tourism, hotel managers are always keen to understand their hotel's perception in the mind of tourists, so that they can take actions accordingly. Whether it's to strengthen hotel's already built-in advantage, or to improve its weakness.

Secondly, the feedback of customers also contains great value, such as to detect the attitude of customer toward the hotel. While it seems not very useful in this case, for the dataset also collects the attitude rating from customers, however, if we can build up a rating prediction model via this dataset, we can apply it to many other scenarios where we don't explicitly receive the attitude scores from customers.

Thus, the expected goals will be as follows:

1. Using LDA to classify each hotel review into one topic(class) and, therefore, assigning the best trait for each hotel based on the most frequent topic reflected in reviews.
2. Conducting sentiment analysis to classify and predict each hotel review into binary positive/negative classes.
3. Taking other external features along with sentiment feature derived from sentiment model to build up a regression model for customer's continuous rating prediction.

[1] examples: EDA analysis, Bag of Words Modeling

Benchmark Model

For LDA model, there will be no benchmark model, it’s rather like an unsupervised learning where topic segmentation and interpretation will be more of a subjective matter.

For rating prediction model, in order to utilize the info from review text, it first takes in the binary output derived from sentiment analysis models as input feature and then combine with other features to train a regression model for continuous rating output.

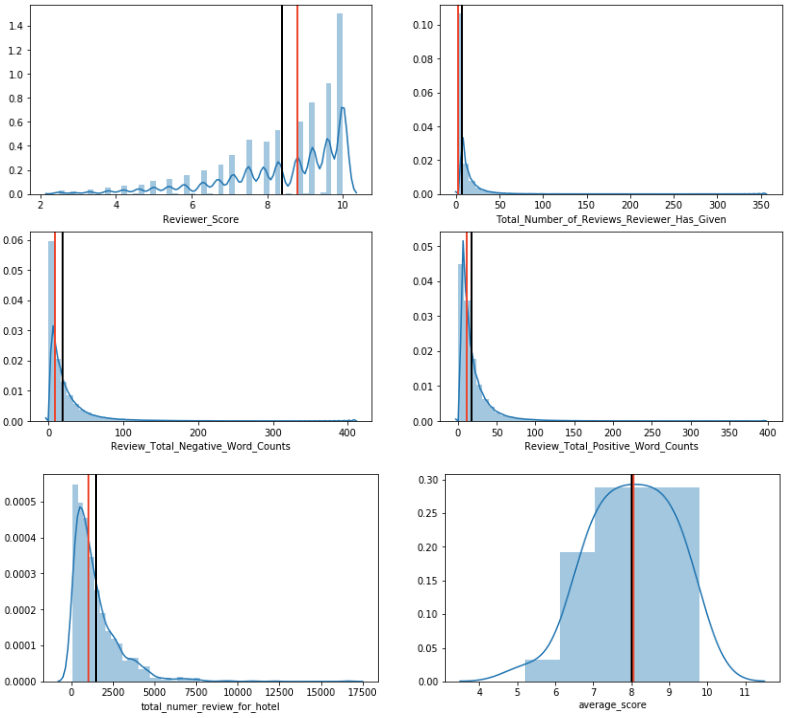
As for now, it seems not having a NLP model being directly applied to predicting continuous output yet. Most focus on binary classification cases. Hence, for text sentiment modeling, I will take **Naive Bayes Model** as the benchmark. Then the sentiment binary output is used along with other feature for continuous output prediction. For the regression task, I will take the **Linear Regression Model** as the benchmark model. Both Naive Bayes and Linear Regression are pretty fundamental models in text and regression cases.

For this dataset specifically, earlier analyst utilizes Naive Bayes and Logistic Regression to predict its sentiment, both seem to have rather good performance with accuracy around 86%-92%. Nonetheless, the prediction on continuous rating hasn’t been experimented yet.

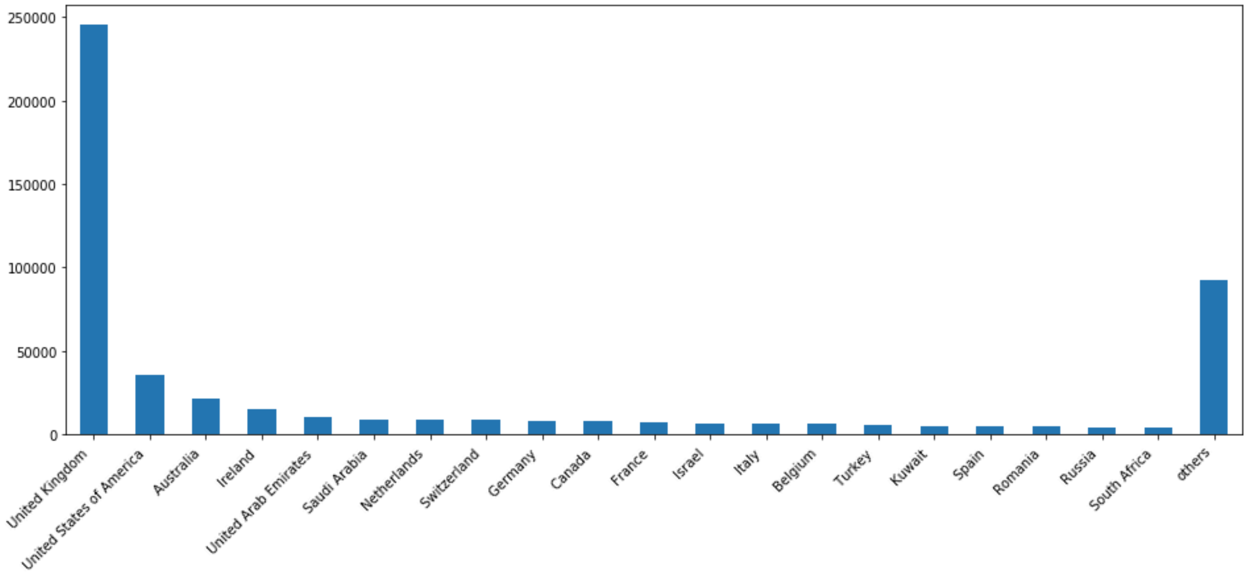
Datasets and Inputs

The dataset contains around 515,000 customer reviews on more than 1,400 hotels. Each review record has 17 features. It is publicly available on Kaggle and was collected by Booking.com. Compared to other review dataset, it contains more information, not only the review text but also the date when attitude rating being made, the reviewer’s nationality and the total reviews hotel receives.

Appendix: Distribution of Numeric Features with Black/Red Line for Mean/Median



Appendix: Frequency Distribution of Reviewers' Nationality



In numeric features, the **Reviewer's score** will be my target variable for both sentiment (after discretization) and regression model. It is a left skewed feature as seen in diagram above (top left). Possible oversampling will be implemented to counteract unbalanced problem in binary model building.

In categorical features, the **Reviewers' nationality** will be further investigated. In addition, we can tell that most reviewers are from UK.

One thing worth noted is that this dataset explicitly separate positive review from negative review, each has its own message box for customer to fill in. For the brevity of analysis, I will combine both of them into single message box in text analysis.

For **the LDA model**, review text will be the only information utilized. Lemmatization and tokenization will be implemented as in natural language processing.

For **the text sentiment model**, it will discretize the customer rating to 0/1 dummy target variable in accordance to classification purpose.

For **the rating prediction model**, in addition to review text, I will implement feature engineering on **reviewer's nationality** and **the date of review made**.

In my first guess, although await for confirmed, the attitude rating distribution would be different among reviewers of different nationality. Maybe reviewer from Japan will be stricter than reviewer from Brazil. Thus, I will calculate the nationality proportion of reviews each hotel has and add it in as an indicator. The other is to take advantage of date info. I will calculate the quarterly average rating for the hotel prior to the date that review is made and add it in as extra feature on the review record. Furthermore, the derived change-rate of attitude rating will also be calculated, which, I believe, will serve as a useful indicator for predicting attitude rating on each review record as well.

Solution Statement

For addressing the problems mentioned above, I expect **the LDA model** will clearly segment different topics and the related words group from the whole review text, and hence assign topic probabilities to each review. Each review record will be labeled with topic of highest probability and hotel will be labeled based on the most frequent topic it receives among all its review records.

Second, I expect **the text sentiment model** will grasp the major tone in the review text and classify review text into positive/negative attitude accurately.

Third, I expect **the rating prediction model** will be able to correctly predict the tourist's attitude rating toward hotel based on the tourist's review text and other provided external features. The predicted outcome will be in continuous scale, where larger number represents more positive attitude.

Evaluation Metrics

In text sentiment model, the **accuracy** is used as the major evaluation metric, while F1 score is also an option if the classes of reviews in dataset are rather unbalanced.

In rating prediction model, the conventional **RMSE (Root Mean Squared Error)** and **R^2** will be used in the assessment.

Project Design

The first phase is raw data inspection. Briefly speaking, steps include removing duplicated records, checking missing values, glancing at the distribution of numeric features with visual assistance. Next is feature engineering, in which I will check if transformation is needed for numeric features and discretize rating into binary classes for later use in text sentiment modeling. Quarterly average rating and change-rate will also be calculated for the use in rating prediction model.

In word tokenization on review text, dataset will be sampled **50%** for training and validation dataset and tf-idf tokens will be pre-configured up to bi-grams for Bag of Words before LDA model is constructed. Insights on topics will therefore be discussed and hotels will be labeled with most frequent topic it receives from its review records, which are already assigned with the most likely topic.

Following touches the first cornerstone on building text sentiment model. I will again utilize the earlier Bag of Words dataset to train the model. Subsequent Naive Bayes/Multi-layer Perceptron model comparison will also be implemented based on validation dataset. The model will serve to predict the binary sentiment for the remaining 50% dataset. I refer and modify this practice from the suggestion on building model with mixed text feature and external features posted on StackExchange as it is quoted:

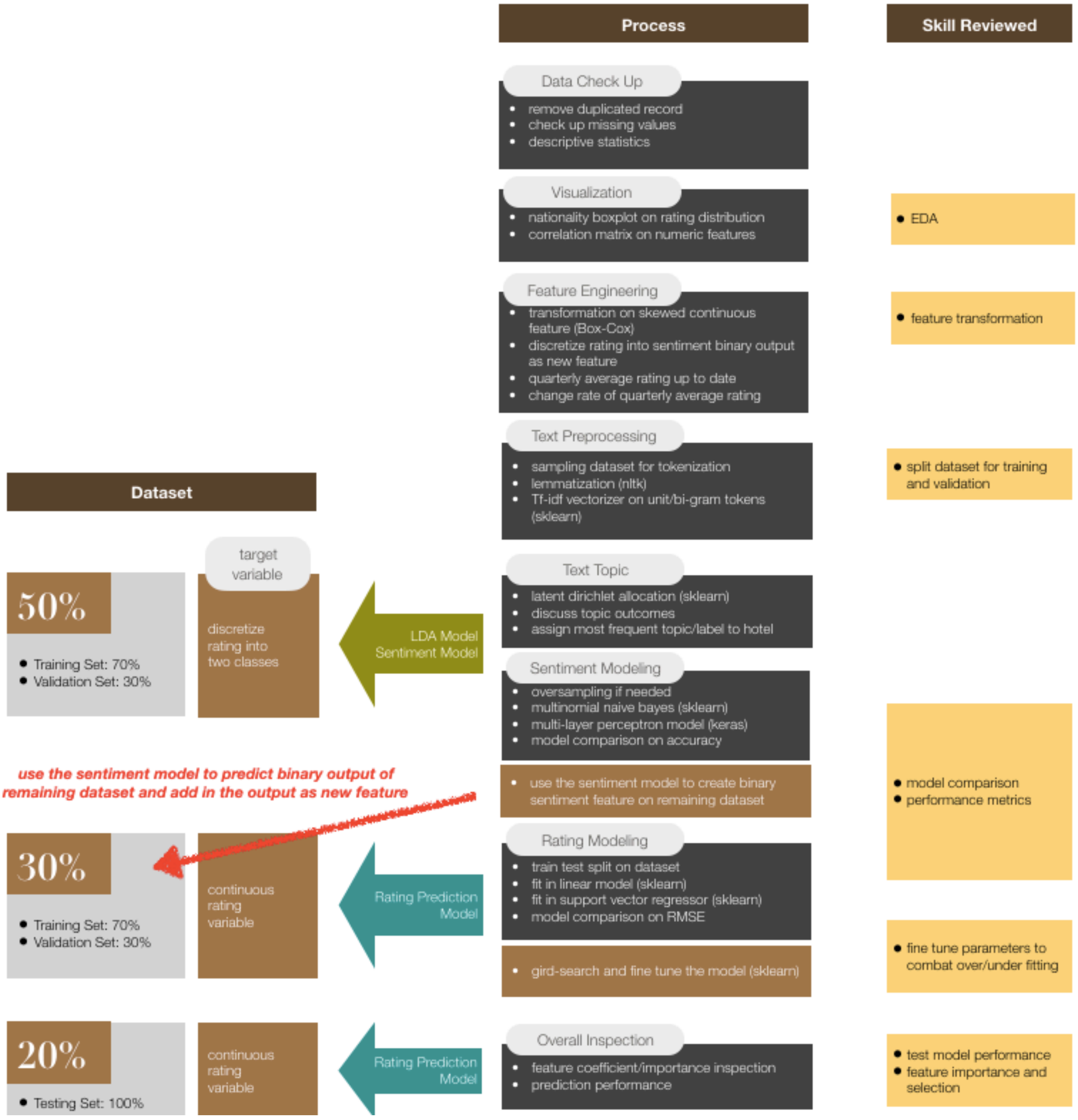
> *Create a model using only your sparse text data and then combine its predictions (probabilities if it's classification) as a dense feature with your other dense features to create a model (ie: ensembling via stacking). If you go this route remember to only use CV predictions as features to train your model*

otherwise you'll likely overfit quite badly (you can make a quite class to do this all within a single Pipeline if desired). by David Jan 6, 2016

Then comes the major task for this project on building rating prediction model. I will again sample 30% of the remaining dataset, which already has the newly add-in binary sentiment output based on prediction from text model, as train/validation data. Again, model comparison will be made on Linear Regression and Support Vector Regressor. Once the final model is chosen, I will conduct the grid-search to fine tune the parameters as the last touch of refinement.

The last 20% dataset will serve as testing data for the final model performance, and brief discussion of feature importance will mark the end of this project.

Appendix: The Project Workflow Diagram



Process

Data Check Up
- remove duplicated record
- check up missing values
- descriptive statistics

Visualization
- nationality boxplot on rating distribution
- correlation matrix on numeric features

Feature Engineering
- transformation on skewed continuous feature (Box-Cox)
- discretize rating into sentiment binary output as new feature
- quarterly average rating up to date
- change rate of quarterly average rating

Text Preprocessing
- sampling dataset for tokenization
- lemmatization (nltk)
- Tf-idf vectorizer on unit/bi-gram tokens (sklearn)

Text Topic
- latent dirichlet allocation (sklearn)
- discuss topic outcomes
- assign most frequent topic/label to hotel

Sentiment Modeling
- oversampling if needed
- multinomial naive bayes (sklearn)
- multi-layer perceptron model (keras)
- model comparison on accuracy

- use the sentiment model to create binary sentiment feature on remaining dataset

Rating Modeling
- train test split on dataset
- fit in linear model (sklearn)
- fit in support vector regressor (sklearn)
- model comparison on RMSE

- gird-search and fine tune the model (sklearn)

Overall Inspection
- feature coefficient/importance inspection
- prediction performance

Skill Reviewed

- EDA
- feature transformation
- split dataset for training and validation
- model comparison
- performance metrics
- fine tune parameters to combat over/under fitting
- test model performance
- feature importance and selection

Dataset

50%
- Training Set: 70%
- Validation Set: 30%

target variable

discretize rating into two classes

LDA Model
Sentiment Model

use the sentiment model to predict binary output of remaining dataset and add in the output as new feature

30%
- Training Set: 70%
- Validation Set: 30%

continuous rating variable

Rating Prediction Model

20%
- Testing Set: 100%

continuous rating variable

Rating Prediction Model