Hotel Data Analysis - Final Report

Zhihan Wang(zw429), Junrui Ye(jy745)

1 Introduction

- Nowadays, a lot of websites are taking advantage of the massive data generated from their users. However, only with the correct methods, can these data best help to understand the users of the site, make prediction of their future behavior and thus help the site generate more profit in the future.

  This is why we would like to see how can the massive data generated everyday benefit Expedia.com, an on-line traveling agent with large scale of services as well as customers. We are using Expedia's hotel-booking data from Kaggle website to analyze the impact of different factors in hotel booking. The detailed goals of this project are to predict whether a hotel will be booked and how much **the gross booking value in USD (GBU**) will be using Julia.

2 Data Cleaning

- **Handling Missing Data:** We select factors that we believe are somehow related to our prediction. Part of the missing values were filtered out except those in GBU column. This is done using Python since the data is too big to process with Julia(all RAM will be used and computer could not handle). So the new data frame contains no NA values in any columns other than gross bookings usd. This gives us a data frame without unnecessary NA values which is small enough for Julia to process yet large enough to build models with without under-fitting problems.

- **Handling Corrupted Data:** To check data corruptions, the columns with Boolean values were checked to see whether the values are all 0s and 1s. For the columns with specific IDs or codes from the website, it is not clear whether the values are corrupted as no information of the detailed codes were given.

3 Data Description

- **Detailed Data Descriptions**

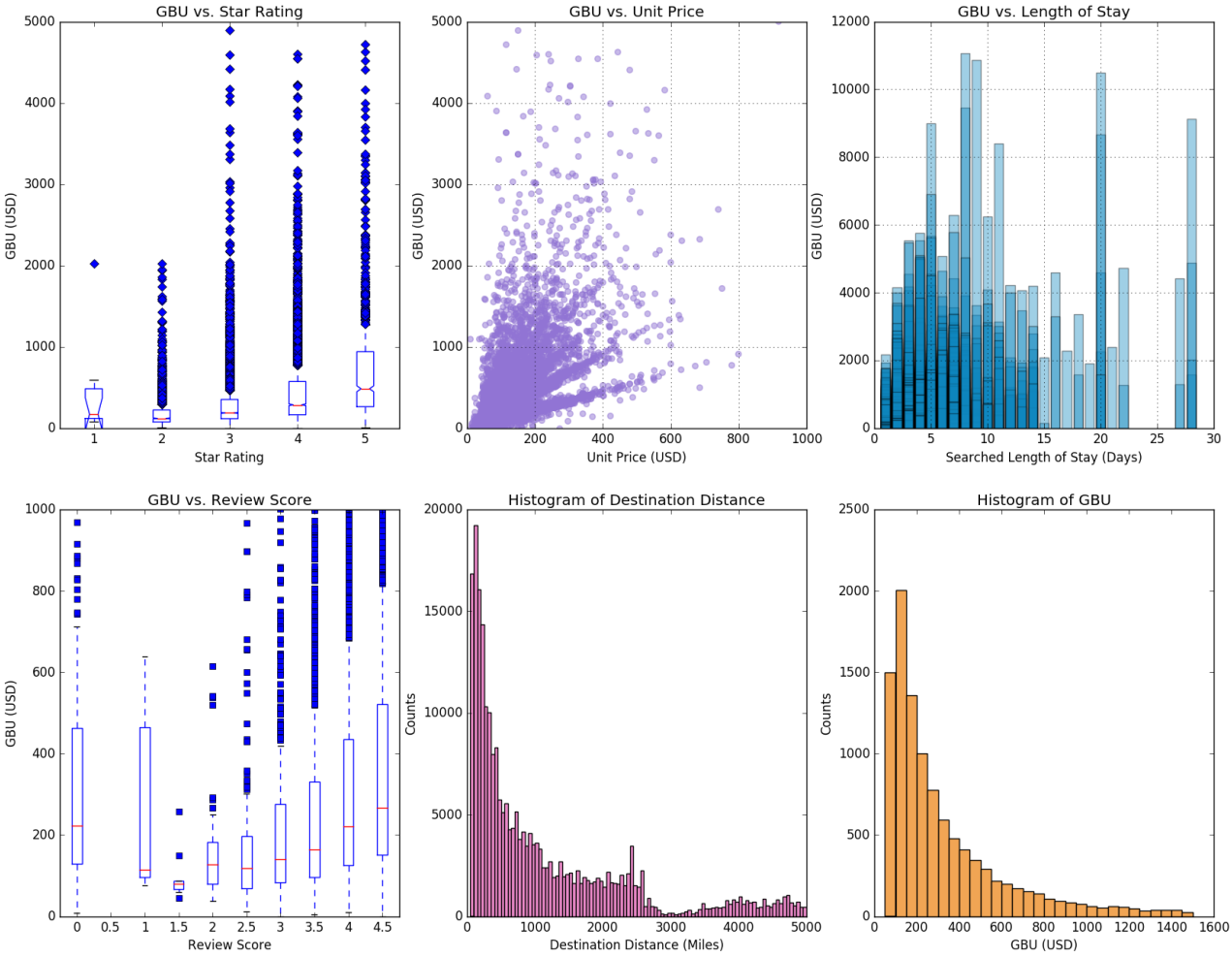
Column Name	Data Type	Description
srch_id	Integer	The ID of the search
date_time	Date/time	Date and time of the search
site_id	Integer	ID of the Expedia point of sale (i.e. Expedia.com, Expedia.co.uk, Expedia.co.jp, ..)
visitor_location_country_id	Integer	The ID of the country the customer is located
visitor_hist_starrating	Float	The mean star rating of hotels the customer has previously purchased; null signifies there is no purchase history on the customer
visitor_hist_adr_usd	Float	The mean price per night (in US$) of the hotels the customer has previously purchased; null signifies there is no purchase history on the customer
prop_country_id	Integer	The ID of the country the hotel is located in
prop_id	Integer	The ID of the hotel
prop_starrating	Integer	The star rating of the hotel, from 1 to 5, in increments of 1. A 0 indicates the property has no stars, the star rating is not known or cannot be publicized.
prop_review_score	Float	The mean customer review score for the hotel on a scale out of 5, rounded to 0.5 increments. A 0 means there have been no reviews, null that the information is not
prop_brand_bool	Integer	+1 if the hotel is part of a major hotel chain; 0 if it is an independent hotel
prop_location_score2	Float	A (second) score outlining the desirability of the hotel's location
prop_log_historical_price	Float	The logarithm of the mean price of the hotel over the last trading period. A 0 will occur if
position	Integer	Hotel position on Expedia's search results page. This is only provided for the training
price_usd	Float	Displayed price of the hotel for the given search. Note that different countries have different conventions regarding displaying taxes and fees and the value may be per night or for the whole stay
promotion_flag	Integer	+1 if the hotel had a sale price promotion specifically displayed
gross_booking_usd	Float	Total value of the transaction. This can differ from the price_usd due to taxes, fees, conventions on multiple day bookings and purchase of a room type other than the one shown in the search
srch_destination_id	Integer	ID of the destination where the hotel search was performed
srch_length_of_stay	Integer	Number of nights stay that was searched
srch_booking_window	Integer	Number of days in the future the hotel stay started from the search date
srch_adults_count	Integer	The number of adults specified in the hotel room
srch_children_count	Integer	The number of (extra occupancy) children specified in the hotel room
srch_room_count	Integer	Number of hotel rooms specified in the search
orig_destination_distance	Float	Physical distance between the hotel and the customer at the time of search. A null means the distance could not be calculated.
price_diff	Float	The diference between price_usd and visitor_hist_adr_usd

- **Main Features to Focus on**

  - The main features to focus on include hotel features like star rating, hotel unit price, review score, customer features like historical booking price of the customer, and search features like length of stay, number of rooms. We choose these features because intuitively they are thought to affect the gross booking values. We perform some preliminary analysis on these data. The relationship between GBU and each of these features were plotted (shown in the next section below) and general patterns were discovered as expected.

4 Descriptive Statistics

- **Plots and Feature Exploration**



- The graphs above show the general correlations of gross booking cost in USD(GBU) with other features specified.
- We have selected star rating, hotel unit price, searched length of stay, review scores to be the main features of the model(discussed in detail in the Analysis Strategies section). This is because it is clear that the first four plots of those features show a great correlation between GBU and the feature specified on the x axis.
- More plots and analysis were also done to evaluate other features which are not shown here. Some histograms of features are presented here to show how the distribution of the features are.

5 Analysis Strategies

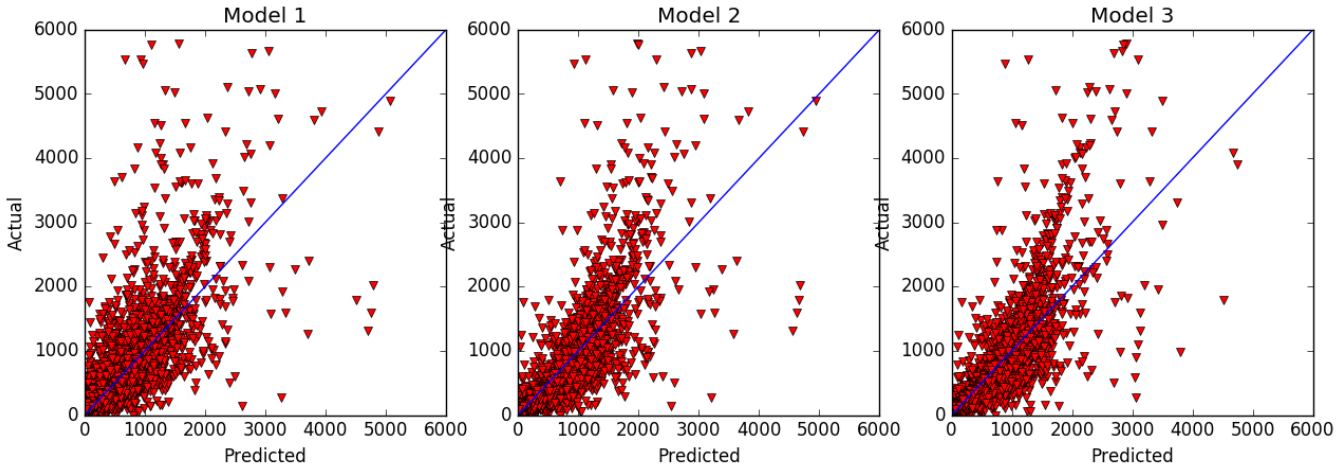
- **Data Manipulation**
  - Since the hotel bookings data is extremely unbalanced (most of the hotel searches do not lead to a booking) and may not be ideal for model training, we randomly select a portion of

no-book entries to create a roughly balanced data set. (37000 datapoints, book vs no-book ≈ 9:11)

- **Feature Engineering**
  - **Feature normalization:** Feature normalization (standard score) was performed on price, price difference and location score to reduce the negative effect of varied ranges of data.
  - **Decision tree:** Decision tree was used for feature selection, because higher positions of the features in the tree imply greater impact of the features in the prediction model. Decision tree was performed on the whole data set first, on which we have found that it is not very insightful since the data set is so unbalanced.
    Thus we have eliminated certain amount of the entries that is a no-book to make the data more balanced. With this new data set, we have found that the most significant features would lead to a booking are: visitor star rating, visitor average booking price, review score, location score, position showed in the web site, price, length of stay, booking window.
  - **Clustering:** Given the fact that customer with different preference may have different probabilities to book different types of hotels, we decided to cluster the customers into several groups based on their past average booking price and star-rating. We also performed clustering on hotels based on features like price, review score and location score. By plotting the clustering results into a density plot shown in the figure below, we can see that, for each hotel cluster, the distribution of the booking rate for each customer cluster is different. We used these clustering labels as features in our prediction model.

Features used for hotel clustering: hotel price, hotel review score, hotel location score.
Features used for customer clustering: customer historical booking price, customer historical booking star rating.

- **Hotel Booking Prediction: Models and Graphs**
  - In order to predict whether a hotel will be booked based on new entries with limited features, several feature engineering methods described above was used to pick the features we want to use in the models. Then, by checking the misclassification rate of the models, we were able to identify the best prediction model with the highest predictive power.
  - **Logistic Regression**: We used logistic regression as our primary prediction model. We believed that logistic regression model can serve our purpose the best because it can generate interpretable coefficients of each feature and probabilistic interpretation of the prediction results. We also used quadratic regularization to prevent over-fitting of the model.
  - Besides logistic regression, we use other models like **Support Vector Machine(Hinge Loss)** and Random Forest to predict the hotel bookings. Detailed analysis of the performance of each model will be shown in the next section.
  - **PCA** was also used for potential improvement of our prediction. However no significant improvement was achieved.
- **GBU Prediction: Models and Graphs**
  - In order to predict GBU, several feature engineering methods described above was used to pick the features we want to use in the models.
  - For preliminary analysis, we fit three different simple ordinary linear regression models:
    Model1: Regression on four features: price, days, review rating, star rating
    Model2: Regression on nine features: price, days, number of rooms, star rating, promotion boolean, brand boolean, location score, booking window, visitor's average booked hotel price
    Model3: Multiply days and number of room together. Other features remain the same as Model2.



- From the preliminary analysis we can see that the features we used have some predictive power, but there were still some patterns that was not captured. Also, outliers are significant in our prediction. So we decided to use more complicated loss function and regularizers to do prediction(See details in result analysis section below).

- **Testing Effectiveness of the Models**

  - Mis-classification Rate
    Compute the mis-classification rate for classification models
  - Mean Square Error
    Compute the mean square error for regression models
  - Cross-validation
    Cross-validation was used in order to check if the models have consistent performance given different training and test sets.

- **Avoiding Under-fitting and Over-fitting**

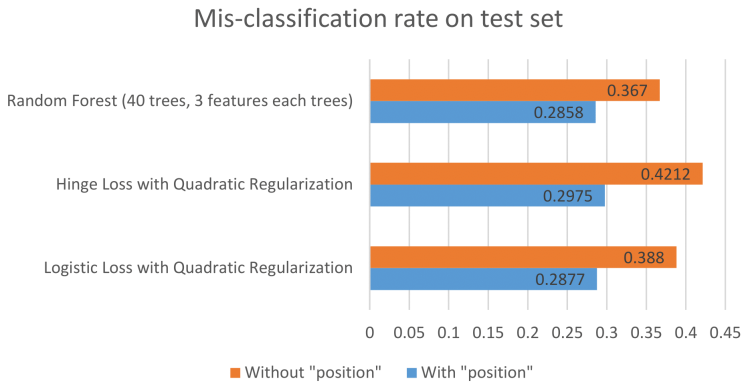
  - To prevent under-fitting, large amount of data was used to perform the analysis on. Compare results from different model in case that certain models do not perform well.
  - To prevent over-fitting, we used regularization to reduce the variance of the model. Also, cross-validation is used to examine whether model have consistent performance.

6 Results Analysis

- **Hotel Bookings Prediction**

  - We found that the "position" feature has significant impact on the accuracy of the prediction. However, the "position" feature was generated by Expedia's hotel ranking algorithm, thus it may carry a lot of other information that we may not know. With this concern in mind, we trained models both with and without the "position" feature so that we can have more meaningful analysis.
  - Features used: visitor star rating, visitor average booking price, review score, location score, (position showed in the web site,) price, length of stay, booking window, brand bool, price diff, room count, customer cluster label, hotel cluster label.
  - Mis-classification rate on test set:
    (Randomly split the data into 80% training set, 20% test set)

Classification Models	With "position"	Without "position"
Logistic Loss with Quadratic Regularization	0.2877	0.3880
Hinge Loss with Quadratic Regularization	0.2975	0.4212
Random Forest (40 trees, 3 features each trees)	0.2858	0.3670



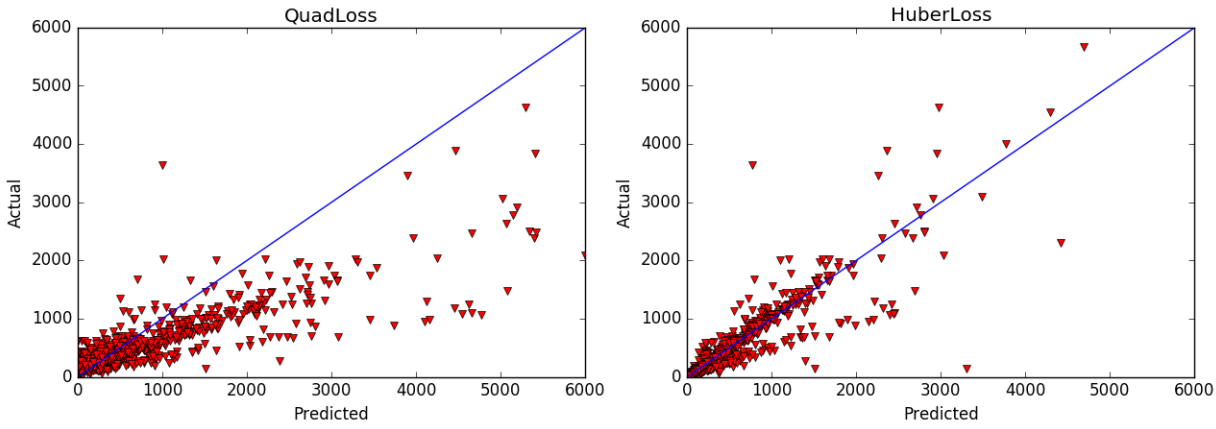
- PCA: no significant improvement after PCA
- From the results above, we can see that Logistic Regression and Random Forest outperform Support Vector Machine in hotel bookings prediction.
- The coefficients of features in Logistic Regression indicates the effect of features on bookings:
  Positive: starrating, review score, room count, location score
  Negative: visitor historical starrating, length of stay, booking window, price, price diff
  Price, room count, starrating have larger impact on the prediction among features used.
- **Cross-validation:** We ran 5-fold cross-validation on Logistic Regression and Random Forest(using features excluding "position"): the average mis-classification rate is 0.3785 for Random Forest, 0.3931 for Logistic Regression. These models achieved similar accuracy in different dataset, so they did not overfit the data.

- **GBU prediction**
  - Features used: price*length of stay*room count, visitor historical average booking price, hotel star rating, review score, location score, booking window, brand bool, promotion bool
  - Results (Randomly split the data into 80% training set, 20% test set):
    QuadLoss V.S. HuberLoss (Quadratic Regularization)
    With the same training set and test set, we can see that HuberLoss significantly outperforms QuadLoss. MSE is $1.1453 * 10^6$ for QuadLoss and 205584.867 for HuberLoss.
  - The coefficients of features in Huber Regression indicates the effect of features on bookings:
    Positive: price*length of stay*room count, visitor historical average booking price, hotel star

rating, review score, location score, booking window, promotion bool

Negative: brand bool

Price*length of stay*room count and location score have larger impact on the prediction among features used.



7 Conclusion

- From the models and visualizations discussed above, we can see that for GBU prediction, Huber loss has provided a rather accurate prediction of the most bookings. However, some systematic outliers still exist, it is believed that such rare data trend is caused by factors like seasonal sales, advanced website membership or occasional discounts from certain hotels.

- We can see that, for hotel booking prediction, random forest and logistic loss with quadratic regularization provided similar prediction accuracy which is better than the result of hinge loss. The result also indicated that the Expedia provided feature "position" plays a rather important role in the data. Without "position" our prediction result is around 10% better than random guessing result and around 30% higher if the feature "position" is utilized in the data models.

- Given the nature of randomness of the hotel booking process, we are somehow satisfied with our prediction results. Nevertheless, there may be more advanced feature engineering and data analysis methods can be implemented to improve the prediction performance.

- Potential usages of our study: We found that location score plays a more important role in boosting hotel booking value than other factors, Expedia can adjust their ranking algorithm to rank hotels with better location score.