Data Mining: Customer Reviews of Hotels

Weikang Wang
dept. electrical engineering
Columbia University
New York, USA
ww2461@columbia.edu

Xinjie Wen
dept. electrical engineering
Columbia University
New York, USA
xw2507@columbia.edu

Peiqi Sun
dept. electrical engineering
Columbia University
New York, USA
email address

***Abstract*—Customer reviews are always important indexes for rating hotels and help them to make progress. In this paper, we compared different models for datasets and compared their performances. Detailed analysis of these methods are also presented.**

***Index Terms*—Support Vector Machine, Naive Bayes, Decision Tree, Data-based Sensitivity Analysis**

I. Introduction

With the help of the development of technology and nearly free and convenient access to surfing the Internet, its becoming common sense to check reviews of all stuff you desire to buy or pay for. When you plan to go out for dinner or purchase a certain new released electronics, its typical of us to look through comments and reviews of those shops or products at the first step and then evaluate whether we are supposed to buy it after deep consideration. Likewise, were getting used to give back our experience, reviews and scores, regardless of positively or negatively, to offer our feedback towards those products. Therefore, its becoming more and more important for companies like hotels to know what kind of features are supposed to pay more attention to for the purpose of gaining more customers and wining rivals in the fierce competitions.

The paper [1] systematically describes the development of methods and tools to analyzing feedback comments of tourism products such hotels and other features which may have an impact on the scores given by customers. And then the author provides a typical flow to utilize data we own, that is extracting useful information from website to get our raw data, cleaning data to filter relevant features and noisy data, feeding preprocessed data into our classification model to predict scores of testing samples, analyzing importance of each features which can be used to provide valuable suggestions for hotels in order to improve their service quality and gain high scores from customers. To conclude, the main goals of this paper are described as follows:

- Building a classification model based on the data fetched from website and features obtained after feature engineering to predict the review scores of testing samples;
- Analyzing the importance and weights of the models features to provide insightful advice to help hotels to improve their service quality.

II. Datasets

Las Vegas, a city born in a desert which is known for gambling and tourism, has started to build plenty of hotels to attract more customers and provide cosy accommodation. Besides, it remains the fast grown large city in the United States over the past ten years. Considering enough useful information like reviews and scores of hotels in Las Vegas, we target Las Vegas hotels to build the classification models. As for features the author has collected, they can roughly be divided into three parts:

- About customers: username, users country, the number of reviews the user has made, the number of reviews about hotels the user has made, helpful votes of the user, member register year, the user continent and the number of member years;
- About reviews: the score of hotels, the review date, the review text, the review language, the period of stay, the traveler type, the review month and the review weekday;
- About hotel: a series of regular facilities, such as pool, gym, tennis court, spa, casino, free Internet. Apart from those, hotel name, hotel starts, and the number of rooms are also taken into consideration.

Surely not all of these features are suitable for constructing a model. Some insignificant features such as users name, review language can be omit. Its obvious that the name of users doesnt account for customers experience. As for review language, since most of the comments are written in English and English is also the most worldwide language, its reasonable not to include that feature. For other unstructure features like review text, it is hard to fit them into a classification model such as support vector machine and these text features need deep learning tools like recurrent neural network to analyze. Therefore the author decided to discard those text features in the classification model.

After collecting data and extracting valuable features, the dataset contained 504 records with 21 extracted features from 21 hotels, 24 per hotel in the year of 2015. The data types of all features include numerical and categorical, which are discrete type and suitable for feeding into support vector machine. Of course, some other types of features can be transformed into categorical type such as Date type.

III. Different models for data mining

A. Support Vector Machine (SVM)

In this project, we used SVM for a regression problem. Support Vector Regression(SVR) is a regression version of SVM, so they share the same principles, with only a few minor differences.

$$minimize||\omega||^2$$

$$subjectto|y_i - (\omega^T x_i + b)| \leq \epsilon_i, \sum \epsilon_i \leq C$$

Followed as the paper, a radial kernel is used to transform inputs into a high dimensional feature space. Since author did not mention the values of parameters in the paper, we used tune function in library e1071 to find the best values for parameters. In our project, cost C is set to be 10 and gamma for radial kernel is set to be 0.01.

B. Data-based Sensitivity Analysis

Sensitivity analysis is a method to extract human understandable knowledge from black box models, like SVM and neural network. Let $\hat{y}$ denote the value predicted by the model for a M-dimensional data sample $x = \{x_1, x_2, ..., x_M\}$ and let P be the function used to calculate output values, so $\hat{y} = P(x)$. In general, the sensitivity methods work by varying an input variable x_a through its range with L levels, under a regular sequence from the minimum to the maximum value. Let x_{a_j} denote the jth level of input x_a. For example, if $L = 5$ and x_a ranges within $[0, 1]$, then $x_{a_j} \in \{0, 0.25, 0.5, 0.75, 1.0\}$. These are the steps of DSA:

- obtain a set of sensitivity responses
  The idea is to apply small changes in one variable of input while fixing the other variables. Problem is that how to set the values of other variables. DSA proposed to randomly pick Ns samples from the original dataset. Then, for each picked sample, x_a is replaced by x_{a_j} and the respective responses are calculated and stored. This procedure is repeated for all input variables, which results in a set of sensitivity responses with the size of $(N_S LM)$.
- VECplot
  VECplot is a plot where x-axis is x_{a_j} and y-axis is corresponding output values. It presents means for visualizing how the output changes with inputs ath feature variable changes. We can draw VECplots by averaging sensitivity responses through Ns different data samples.
- Measure of input importance Following are 4 sensitivity measures to quantify the influence of each feature has on the output value:

$$S_r = max(\hat{y}_{a_j} : j \in \{1, 2, ..., L\})$$

$$-min(\hat{y}_{a_j} : j \in \{1, 2, ..., L\})$$

$$S_g = \Sigma_{j=2}^{L} |\hat{y}_{a_j} - \hat{y}_{a_{j-1}}|/(L-1)$$

$$S_v = \Sigma_{j=1}^{L} (\hat{y}_{a_j} - \overline{y}_a)^2/(L-1)$$

$$S_a = \Sigma_{j=1}^{L} |\hat{y}_{a_j} - \widetilde{y}_a|/L$$

where $\overline{y}_a$ is mean and $\widetilde{y}_a$ is the median. Gradient requires the order of sensitivity responses, which means this measure is not suitable for data with nominal input variables. Range and variance measures are sensitive to outliers. Hence, Average Absolute Deviation (AAD) measure was used in this project.
Let S_a denote the sensitivity measure for x_a. If S_a is small, then ath variable of input is not supposed to be important in this model, because the change of x_a almost does not affect the output. Therefore, the higher the sensitivity measure, the more relevant is the feature variable. We can get relative importance (Importance value) as

$$r_a = S_a / \Sigma_{i=1}^{M} S_i$$

And here is the algorithm for computing the importance value:

Algorithm 1 Calculating Importance Values

```
Sample N_s data points to create SA dataset
for each data point x in SA dataset do
  for a ∈ {1, ..., M} do
    for j ∈ {1, ..., L} do
      Feed the new input x' such that x'_i = x_i if i = a and = x_{a_j} if i ≠ a
      And get corresponding ŷ_{a_j}
    end for
  end for
end for
To this end, an (N_s * M * L) Output matrix Ŷ is obtained.
Average Ŷ though its first dimension to get a M * L matrix Ŷ_ave.
for a = 1, 2, ..., M do
  Plot Ŷ_ave[a, :], which is a vecplot for x_a.
end for
for each variable do
  compute ADD measure: S_a = Σ_{j=1}^{L} |ŷ_{a_j} - ỹ_a|/L
  Compute Importance Value: r_a = S_a / Σ_{i=1}^{M} S_i
end for
```

C. Decision Tree

Decision Tree is a simple but useful method for data with high dimension. For our dataset, every data point has 19 dimensional feautures and it's hard to adopt other low-dimensional models to fit it, such as linear regression. For decision tree model, one of the biggest advantage is that it can show clearly what features influence mostly the decision of the prediction procedure and the threshold for each node.

For the customer review dataset, there are 504 data points and each with 19 dimensional features. By using regression decision tree method with pruning technique, we hope that this model could have more generation ability than the SVM model. We show that result in the following section.

D. Naive Bayes Classifier

Naive Bayes Classifier is one famous generative classifier that put priors for each class density distribution and assume each feature are independent given class label. By using generative classifier, we could deal with the unbalanced data problem.

In our paper, we assume that the class density distribution of each feature under each class is gaussian distribution and this is just the mixture of gaussian model. We show the result of fitting the dataset in the following section.

IV. Experiments and result analyzing

We adopted support vector machine and data-based sensitivity analysis which are used in the original paper and reproduced the result. We also used decision tree and naive bayes classifier on the customer reviews datasets and analyzed their results and performances.

A. Reproduction of the original paper

A.1 Reproduction of the SVM result

We assessed the predictive performance of our model by running a 10-fold cross-validation 20 times, as the same as the paper described. Performance metrics used here are MAE and MAPE. MAE is the mean absolute error between predict scores and real scores. MAPE is the mean absolute percentage error. For example, if real score is 4 and predict score is 5, then absolute error is 1 and absolute percentage error is one fourth. For both metrics, the lower the value, the better is the modeling performance.

The results from different models used in our project can be seen on Table I . Compared to the result from the paper, our reproduced SVM achieved a bit higher MAE and MAPE, so it performs almost as good as the original model.

Since we discard the review text information, it is hard to fit a super accurate model. However, other studies proved that a model with a MAPE of around 27% is accurate enough to extract some valid insightful knowledge, we believe some useful knowledge can be extracted from these models.

TABLE I
SVM Results comparison

Classifiers	MAE	MAPE
Original Paper Result	0.745	27.32%
Our Result	0.764	27.85%

We shows similar patterns in Figure 1 and Figure 2, indicating that our SVM model is similar to the original SVM model in the paper. But this also shows that SVM models have a poor prediction performance for instances with a small real score, like 1 or 2. The reason for this is that our datasets is unbalanced, composed of only 41 instances with a real score less or equal than 2 and 463 instances with a score bigger or equal than 3. Hence, the SVM model trained from this datasets is biased.

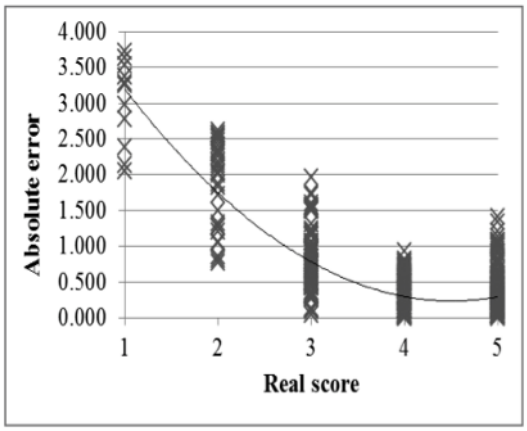

Fig. 1. Scatterplot of real scores versus absolute error from the original paper.

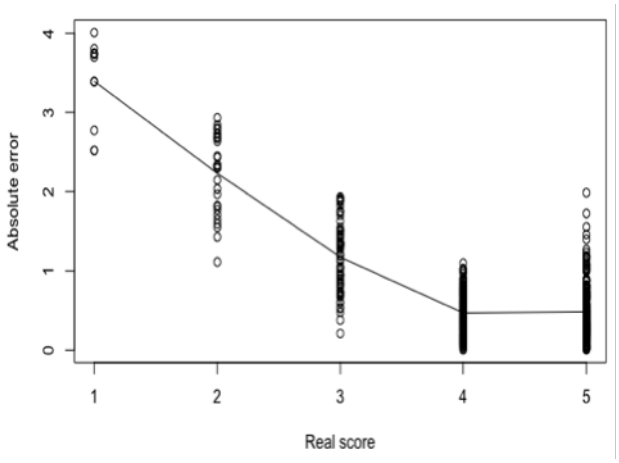

Fig. 2. Scatterplot of real scores versus absolute error of our model.

In figure 3, we can see that our model plays well when the true score of the original data is high (bigger than 3), while plays poorly when score is low (less than 3). This happens because we has few training data points of low score. We can find that naive bayes classifiers and decision tree methods behaves better in the following.

A.2 Reproduction of the Data-based Sensitivity Analysis (DSA)

As mentioned before, DSA is applied to analyze relationships between output value and input features based on a black-box regression model, which is SVM in this project. Figure 4 is a bar graph exhibiting the percentage relevance for all features from the paper and Table II is a table about the percentage relevance for all features calculated from our SVM model. As we see from Figure 4 and Table II:

- Nr. Hotel reviews, Member years, Nr. Reviews, Nr. Rooms, Hotel Stars are the top 7 most relevant features in both 2 models.

	pred				
true	1	2	3	4	5
1	0	0	1	2	8
2	0	1	3	8	18
3	0	0	16	20	36
4	0	0	1	84	79
5	0	0	2	13	212

Fig. 3. True-pred Table of our SVM

- The existence of Gym, Spa, Casino, Tennis court and Pool almost does not affect predicted Scores for hotels. The reason can be that most hotel costumers only stay in their rooms instead of having fun in swimming pool, casino or Spa.
- Free internet plays a role in our model with a relative importance of 6 %, but is not important in the papers model with a relative importance of 2.7%. On the other hand, Period of stay plays a role in the papers model with a relative importance of 10.3 %, but is not important in our model with a relative importance of 2.7%.

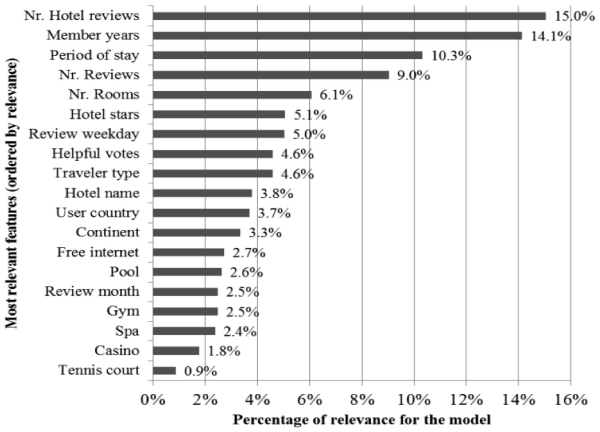

Fig. 4. Most relevant features in the paper

A.2.1 Vecplots for Nr. Hotel reviews and Nr. Reviews

Nr. Hotel reviews is the total number of hotel reviews the user has published on tripadviser while Nr. reviews is the number of all reviews including not only hotel, but also restaurants, attraction units and so on. Hence, these 2 features are correlated with each other, and they should have similar impacts on the predicted score. As stated previously, Nr. Reviews has a relative importance of 9% in the paper and 16% in our model. Nr. Hotel reviews has a relative importance of 15% in the paper and 7% in our model. These 2 features are important in both 2 models.

TABLE II
MOST RELEVANT FEATURES IN OUR MODEL

Classifiers	MAE	MAPE	
Feature name	Relevance	Feature name	Relevance
Member years	14%	Pool	3%
Hotel stars	10%	Traveler Type	3%
Nr. Rooms	8%	Review month	2%
Nr. Hotel reviews	7%	Period of Stay	1%
Free Internet	6%	Gym	1%
Helpful Votes	6%	Tennis Court	1%
User Country	6%	SPA	1%
User continent	5%	Casino	1%
Review Weekday	4%		

Figure 5, Figure 6 and Figure 7 show that, in the original model, as the number of Hotel reviews increases, predicted score will first decrease from 4.4 then converges to 4.1; as the number of reviews increases, predicted score will also first decrease from 4.4 then converges to 4.0. On the other hand, in our result, as the number of Hotel reviews increases, predicted score keeps decreasing from 4.4 to 3.8; as the number of reviews increases, predicted score keeps decreasing from 4.4 to 2.6. There is a little difference between our model and authors model, since parameters can be set different. However, the trend of these plots are almost same and these results are also aligned with studies that claims the initial first experiences tend to turn the customer more demanding when publishing online score and so global reviews may have the effect of plunging scores to values below 3.9.

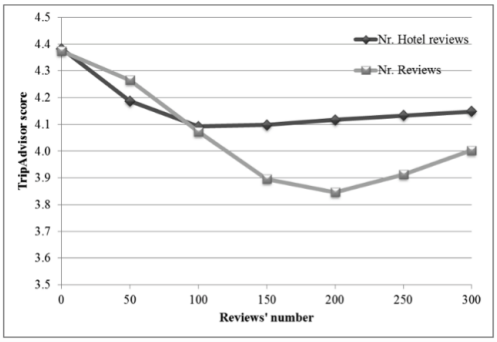

Fig. 5. Influence of "Nr. Hotel reviews" and "Nr. Reviews" on score in the original model

A.2.2 Vecplots for Nr. Rooms

Nr. Rooms indicates the total number of rooms the hotel has. As stated previously, Nr. Rooms has a relative importance of 5% in the paper and 8% in our model. Thus, this feature is important in both 2 models.

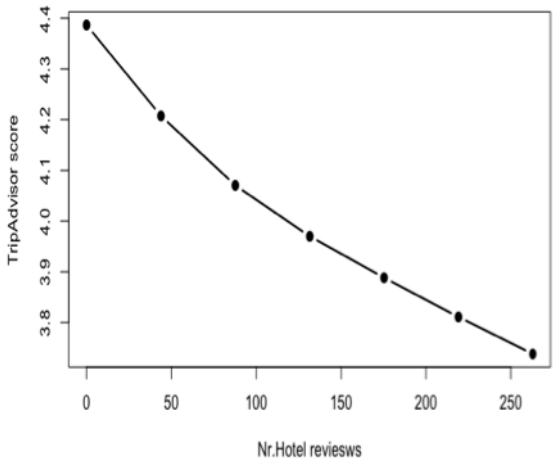

Fig. 6. "Nr. Hotel reviews" versus score in our model

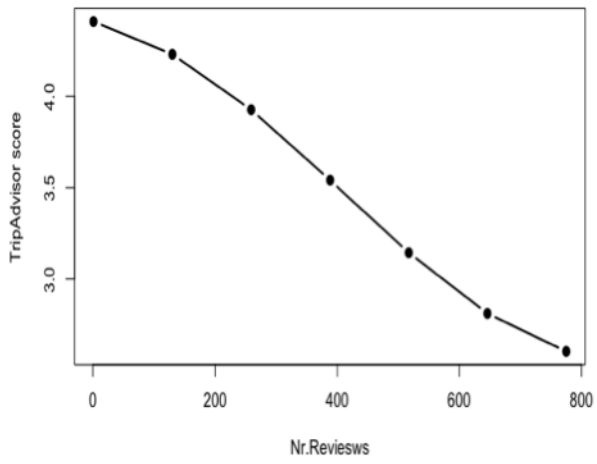

Fig. 7. "Nr. Reviews" versus score in our model

As we can see from Figure 8 and Figure 9, our model gave a vecplot for Nr. Rooms similar to the papers. In both 2 plots, as the Nr. Rooms increases, predicted score decreases from 4.5 (4.65) to 4.1. This finding suggests that hotels with too many rooms, which means a lot of costumers, may fail to offer a high-quality service for all costumers. Hence, we can know that small hotels have an advantage on giving costumers a friendly and non-crowded environment.

A.2.3 Vecplots for Nr. Stars

The hotel star rating is a widely-used measurement for hotels quality. Of course this feature should be relevant with online rating score. As stated previously, Nr. Stars has a relative importance of 5.1% in the paper and 10% in our model.
Figure 10 and Figure 11 show that, as the value of hotel star rating increases, predicted score in the

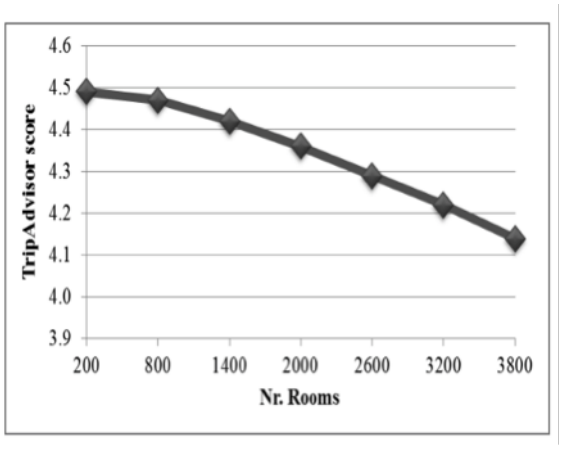

Fig. 8. "Nr. Rooms " versus score in original model

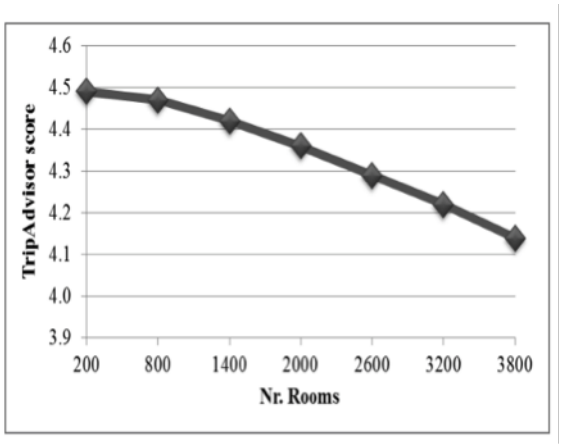

Fig. 9. "Nr. Rooms" versus score in our model

paper increases from 4.15 to 4.5, while predicted score in our result increases from 3.9 to 4.7. It seems that our model is more dependent on this feature.

A.2.4 Vecplots for Weekday

As stated previously, Weekday has a relative importance of 5.1% in the paper and 4% in our model. Although the relevance of Weekday is small
In figure 12, the x-axis (weekday) is not sorted in a way that we are familiar with, since when R plots it, the weekday variable is sorted in strings mode. Still, if we list the weekday variable in descending order of its corresponding predicted value, we will get same result: Saturday, Friday, Sunday, Thursday, Monday, Wednesday, then Tuesday. Hence, we believe review weekday has an influence on costumers rating. For example, people who come on Friday or Saturday are happier and more positive, since they can enjoy the weekend time; people who come on Tuesday could be still at work in hotel.

A.2.5 Vecplots for Period of Stay

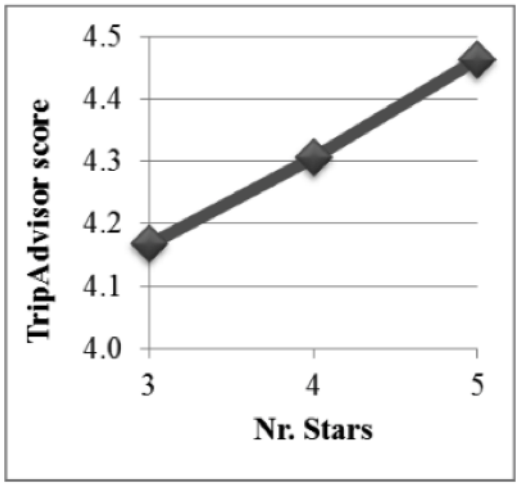

Fig. 10. "Nr. Stars " versus score in original model

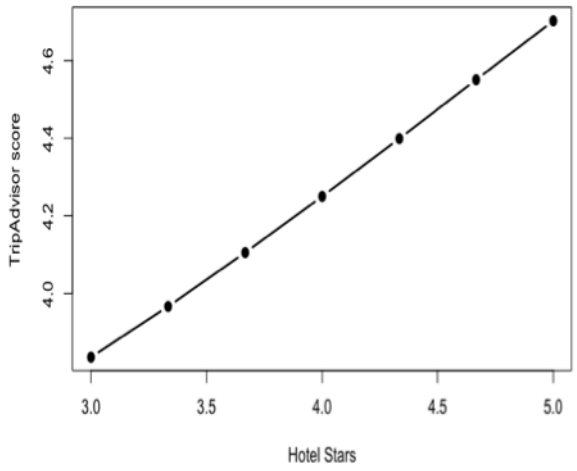

Fig. 11. "Nr. Stars" versus score in our model

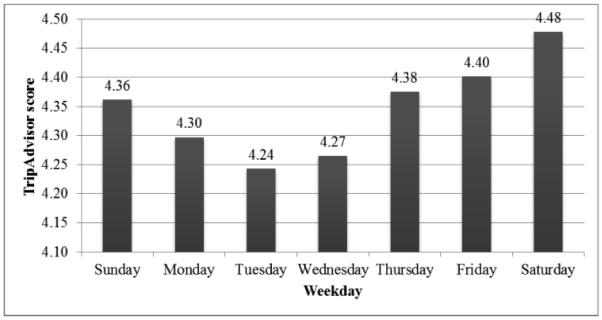

Fig. 12. "Weekday " versus score in original model

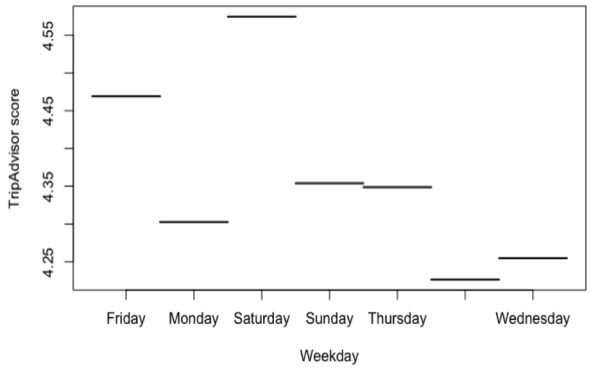

Fig. 13. "Weekady" versus score in our model

Period of Stay refers to in which season that the costumer came. In the paper, Period of Stay with 10.3% relevance is considered to have a seasonality effect on rating score, because Las Vegas is a desert city and more attractive in colder season. However, the range of scores with different seasons is tiny, only 0.07, so it looks like a result of noise, which is supported by our result. This is also a warning for us that due to the small dataset, our findings are not strong enough to be one hundred percent true.

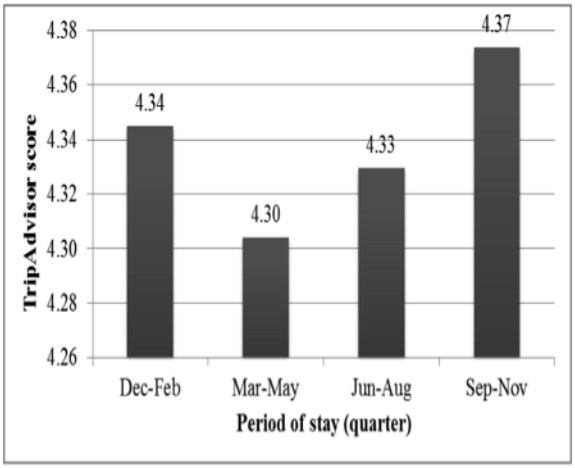

Fig. 14. "Nr. Stars " versus score in original model

A.2.6 Vecplots for free internet

Author did not provide a vecplot for free Internet, because of its low relevance in his model, only 2.7%. But in our model, free Internet has a relevance of 6% and so is considered as an important input feature. As we can see from figure 9, hotels with free internet are predicted with a score 0.25 higher than those without free internet. This is aligned with our expectation, because at the present time, people cannot live without internet.

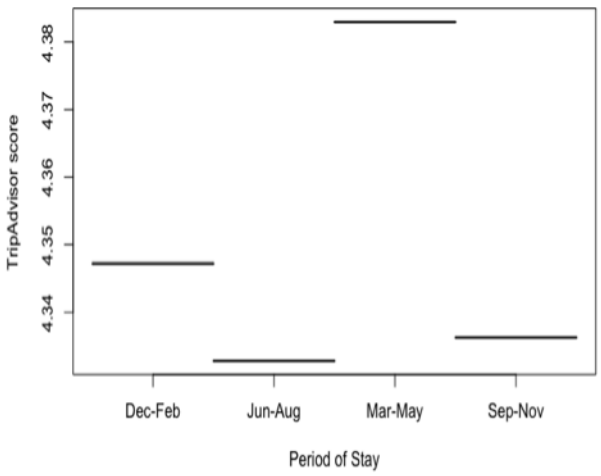

Fig. 15. "Nr. Stars" versus score in our model

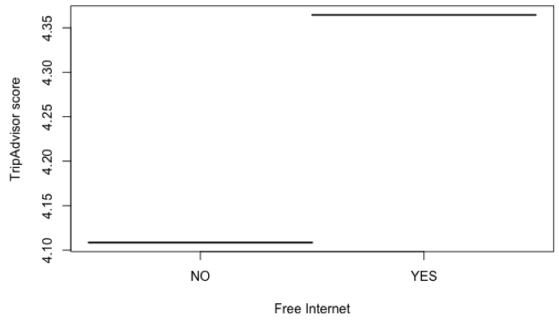

Fig. 16. "Free Internet" versus score in our model

B. Results of the Decision Tree

We also used the decision tree method for the customer review datasets. Data points from the original datasets have 19 dimensions and by constructing a decision tree for this dataset, it could easily get overfitting. Thus we adopt tree pruning method. The pruning tree is shown in Figure 17:

We can see from the resulting tree that, it only uses three features of the original data points. This shows that the decision tree is more robust and simpler than the SVM mothed. By the way, we can also see that these three features correspond to the important features selected by the Data-based Sensitivity Analysis, which is a connection between this two methods.

In Table III, we present the MAE and MAPE of the decision tree model. We can see that these two numbers are lower than SVM's, indicating that decision tree has more generation ability than SVM for this dataset, i.e., decision tree is a more robust and proper model for this dataset.

C. Results of the Naive Bayes Classifier

We also adopt the Naive Bayes Classifier model on the original datasets. Naive Bayes Classifier is a generative model, which is different from Decision Tree and SVM models,

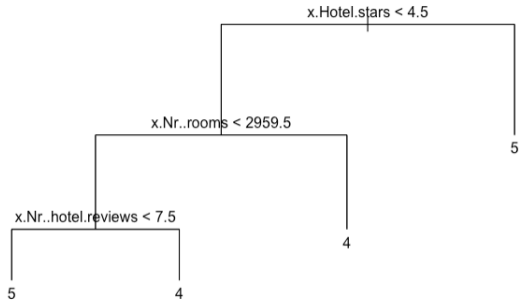

Fig. 17. Decision tree for the original data

TABLE III
DECISION TREE RESULT

Classifiers	MAE	MAPE
Decision Tree	0.684	27.66%
SVM	0.764	27.85%

i.e., discriminative models. For generative models, the most significant advantage is that that can deal with the problem that some kinds of data points are much fewer than others. In our dataset, data points with ture score 1 and 2 are much fewer than the data points with ture score 4 and 5. So we hope that Naive Bayes Classifier methods can over perform than SVM and Decision Tree methods.

In our model, we assume that each class density distribution is Gaussian and thus we get a mixture gaussian model. We show the result as in Figure 18 and Figure 19:

	pred				
ture	1	2	3	4	5
1	5	0	0	0	1
2	0	7	0	0	0
3	1	1	17	5	3
4	2	4	4	51	18
5	5	3	5	8	60

Fig. 18. Naive Bayes Classifier for training data

We use 200 data points for training and 304 data points for testing. The MAE and MAPE for Naive Bayes Classifier are shown in Table IV.

TABLE IV
DECISION TREE RESULT

Classifiers	MAE	MAPE
Decision Tree	0.684	27.66%
SVM	0.764	27.85%
Naive Bayes Classifier	1.21	33.77%

We can see from the result that the Naive Bayes Classifier

	pred				
ture	1	2	3	4	5
1	0	1	0	1	3
2	2	2	2	7	10
3	5	0	3	17	20
4	4	3	8	35	35
5	8	6	6	52	74

Fig. 19. Naive Bayes Classifier for the testing data

does not result in better performance in regrad of MAE or MAPE. Actually, it does worse than the other two in these two measurements. But for another measure, we can find that in both training and testing result, the Naive Bayes Classifier both give more separated prediction labels for data points with any scores. Thus it does not neglect the ones with true score 1 and 2, and partially solves the problem that there are few data points with score 1 and 2 by introducing a prior for each class density distribuiton.

V. Conclusion

The dimension of our data was 19 and there are only 504 data point, thus its hard to fit this dataset. By using three different models to fit the data, we can find that they have different results for the same dataset.

For the SVM classifier, it uses all features as a whole vector and use radial kernel. From the result we can see that the data was well separated in this kernel since the MAE and MAPE was low.

For decision tree, it just uses some important features of the data and the chosen features corresponds to the important features of the data selected by DSA.

For naive bayes, since its a generative model, it has priors for each class. So it can handle the problem that we only has few data points for score 1 and 2 situations. We can see from the result that it truly deals better with low score data than the other two models.

Thus for a dataset, since we do not know the original generating model, there is generally no a best model for this dataset. Each model may have its own advantages over others. In real scene, we need to use cross validation method or bayesian model selection method or other modern methods to find a proper model for our dataset.

References

[1] Srgio Moro, Paulo Rita, and Joana Coelho. Stripping customers' feedback on hotels through data mining: The case of las vegas strip. *Tourism Management Perspectives*, 23:41–52, 2017.