The Big Data Analytics Project Report

Prediction on Yelp dataset review rating

Heng Lyu

501064765

Supervisor: Dr. Ceni Babaoglu

2021-12-03

Table of Contents

Abstract

While traditional data management has its limits with the large amount, quick pace and big diversity data in the society nowadays. Artificial intelligence, internets also mobile phones generate a lot of data with fast speed. The development of the big data analytics is significant for the rapid decision creation, shaping and forecasting expected future results and improved intelligence in some industries. The advantage of big data analytics is obvious, which includes the low cost, high flexibility and big capacity. Yelp has an incredible number of users' behavioural data as the leading company in online advertising industry, and this project will dig deeply into the Yelp's open dataset to find information behind reviews and build a classification model to predict review ratings on restaurants in British Columbia, Canada.

Introduction

The label of model will be the star ratings in each unique review which predicted by businesses and users' features such as business attributes, business categories, opening hours, users' review counts, and so on. In terms of model outcomes, this would be an obvious supervised classification question. This project will try different classification methods, such as Naïve Bayes, KNN, Random Forest and LightGBM. The review sentences in review.json was dropped because they are difficult to analyze without using sentiment analysis. The main purpose of the project will concentrate on predict rating scores based on provided categorical and numeric features. The research questions are as follows:

1. Which are the top 10 restaurants in British Columbia, Canada?
2. What factors of restaurant and users will influence users' rating in reviews most?
3. Which classification model will predict review rating most accurately?

Literature reviews

Since the Yelp dataset is famous for education use, there are lots of articles about this dataset. We decide to review literatures through 3 aspects. Firstly, we will review researches on Yelp dataset from different fields. Secondly, we will dive deep into Yelp dataset review rating prediction, and discuss about similarities and differences between this project and other articles. Third, we will explore more about rating prediction for other datasets, and discuss what we can learn from other studies.

General study on Yelp dataset

Exploratory Data Analysis and Data Mining on Yelp Restaurant Review (E. S. Alamoudi and S. A. Azwari 2021)

In the paper "Exploratory Data Analysis and Data Mining on Yelp Restaurant Review," authors state that EDA is the useful Methodology for restaurant rating system in the data mining projects. By focusing on the fast-food chains, like KFC and McDonald's, the Yelp dataset has offer volume data for analyzing the features of the restaurant reviews. Using Term frequency method to identify the customer experience and get the prediction of the review results. By observing various data, we can see the frequent words, different scoring levels, and location characteristics of the scoring system. However, the Bag-of-Words model and N-grams are both for Natural Language Processing which will not be mentioned in our project. The data mining and analysis process helps to develop the restaurant business. It suggested to use Plotly and Cufflinks packages in python for data visualization.

The cultural impact on social commerce: A sentiment analysis on Yelp ethnic restaurant reviews (Nakayama, M., & Wan, Y. 2018)

Why is digital review data analysis important for restaurants? For example, the authors Zhongshan Cheng and Yun Wan have figured out the cultural effects of social commerce in the catering industry. They use Yelp data, especially in the reviews of Japanese ethnic restaurants, to see how different cultural backgrounds have different evaluations of food quality, service, atmosphere, and price fairness. Through some valid hypothesis, after data mining and analysis, the preferences of individuals from different cultural backgrounds can be obtained.

Predicting the Helpfulness of Online Restaurant Reviews Using Different Machine Learning Algorithms: A Case Study of Yelp (Luo, Y., & Xu, X. 2019)

There are many machine learning algorithms suitable for analyzing the results of restaurant review systems. In the “Predicting the Helpfulness of Online Restaurant Reviews Using Different Machine Learning Algorithms: A Case Study of Yelp,” authors use the Latent Dirichlet Allocation (LDA), Support Vector Machine (SVM), Fuzzy Domain Ontology (FDO) algorithms, Naive Bayes (MB) and SVM ontology to find sustainable utilization results in the industry. And how it can affect the country’s sustainable economic growth. By using different machine learning algorithms to analyze 294,034 comments on Yelp.com, the data mining results will help restaurants satisfy customers and contribute to sustainable economic growth. It indicates that positive reviews mentioned food tastes a lot, and negative reviews focused more on value. However, there does not exist relevant features in our business dataset to use in this project.

Categorizing health-related cues to action: using Yelp reviews of restaurants in Hawaii (Ariyasriwatana, W., Buente, W., Oshiro, M., & Streveler, D. 2014)

One of the hot topics in the restaurant industry is about the food quality and the health issue. Writers in the "Categorizing health-related cues to action: using Yelp reviews of restaurants in Hawaii" has observed data in Hawaii to explain the importance of the health-related commentaries in the restaurants. And concluded that health-related commentaries need the restaurant reactions based on the datum analysis. Since authors pay attention to the health-related issue. The data mining project sets the categories of reviews based on the health-related features as well. Such as nutrition etc... While few restaurants in Hawaii, the categories and coding scheme are easier to build up. The authors use two different coders in the paper to complete the examination of the health-related reviews and see how the actions required to have the better feedback from customers.

Modern Food Foraging Patterns: Geography and Cuisine Choices of Restaurant Patrons on Yelp (Q. Xuan, M. Zhou, Z. Zhang, C. Fu, Y. Xiang, Z. Wu, & V. Filkov. 2018)

Some people think that geography is an important model for determining food preferences. Therefore, some authors have conducted some research on how geography affects food choices. For instance, authors in "Modern Food Foraging Patterns: Geography and Cuisine Choices of Restaurant Patrons on Yelp," setup their data with customer reviews in different locations to see whether people have the geography preference in choosing restaurants. In the paper, TSN, GFN, MLR and TFN are the methodologies used to discovery the customer’s behaviors. By discovering the rules of customer behavior from data, machine learning algorithms are applied to design recommendation systems and further research is carried out.

Yelp review rating prediction

Characterizing non-chain restaurants’ Yelp star-ratings: Generalizable findings from a representative sample of Yelp reviews (Keller, D., & Kostromitina, M. 2020)

This article focused on extracting features from review text for non-chain restaurants. The authors tried to figure out criteria which customers use when rating the restaurants. They applied Multiple Correspondence Analysis to reduce dimensions of corpus which collected from Yelp recent restaurant reviews. Although our dataset has dropped the review comment column, we still can try this method or similar methods for reducing dimensions when analysing business categories. It is important to notice that this article only concentrate on non-chain restaurants; in contrast, both chain restaurants and non-chain restaurants are mentioned in our project. The authors conclude that service quality has the most impact on 1-2 stars reviews, and food quality has the most impact on 3-4 stars reviews. It implies that there may exist a strong correlation between review ratings and business attribute in the future model in our Project.

Yelp Review Rating Prediction: Machine Learning and Deep Learning Models (Liu, Z. 2020)

This article is quite similar to our project. The source data is composed of review.json and business.json in the article. Meanwhile, our dataset extracts columns from review.json, user.json and business.json. Also, there are many machine learning methods mentioned in this article such as Naïve Bayes, Logistic Regression, Random Forest and Linear support vector machine. The article explored more on Transformer-Based Model which is an important model in Natural Language Processing field. Since our dataset dropped review text column, we will not use any method related to NLP. However, our project will try more machine learning models especially in decision tree. Our project will not only apply random forest, but also try algorithms like xgboost, or light gbm. In addition, our project will try to optimize parameters in tree models manually.

Robust Review Rating Prediction Model based on Machine and Deep Learning: Yelp Dataset (A. Rafay, M. Suleman and A. Alim. 2020)

Similar to last article reviewed, this article predicts review ratings through NLP and sentiment analysis as well. The author mainly used Multinominal Naïve Bayes as machine learning model and Convolutional Long Short Term Memory as deep learning model. However, our project will not involve these 2 models since review sentences are removed from dataset. Although sentiment analysis will not be conducted in our project, this article still have some points to notice. It indicates that it is necessary to apply cross-validation approach to dataset in order to overcome high-biased dataset. As a result, each folder should contain almost equal percentage of 5 labels from 1-star to 5-star.

Rating prediction on other datasets

Enhanced review-based rating prediction by exploiting aside information and user influence (Wu, S., Zhang, Y., Zhang, W., Bian, K., & Cui, B. 2021)

The Yelp reviews are the historical data that generated from the customers in the market. There are valuable comments in the database, but there are also invalid comments. The use of ERP helps to evaluate the effective and supportive comments in the data set. Only the efficacious and highly topic related reviews can provide support for the decision-making process. The advantage of ERP lies in conforming to the scenarios in the comment system and observing influential comments. Therefore, by observing all features, we can consider rating predictions and optimal decisions. Authors in the "Enhanced review-based rating prediction by exploiting aside information and user influence" believe ERP can help its users to get the item-aware preference and improve the efficiency of result prediction

User-Personalized Review Rating Prediction Method Based on Review Text Content and User-Item Rating Matrix (Wang, B., Chen, B., Li, M., & Zhou, G. 2019)

According to Wang, B., Chen, B., Li, M., & Zhou, G, the current rating prediction has the disadvantage on separating different users. The personalized and specific reviews required the attentions on the user's special characteristics, and it will improve the prediction system. The author in the article uses the UPRRP method to focus on user reviews and rating matrices. And based on the observed data set, the UPRRP model has more advantages over the RRP method in personalized results. The RRP method is based on Support Vector Machine, and the UPRRP method apply a linear regression after RRP to predict rating. Although the features extracted from review text will not be mentioned in our project, it is still necessary to apply feature engineering on user data.

Predicting the ratings of Amazon products using Big Data (Woo, J., & Mishra, M. 2020)

This paper used Amazon products dataset which contains 15 attributes and has about 7 million records. The rating star column was used as label column, which means the Amazon products dataset is similar to our Yelp dataset. Since the dataset is really large, the author decides to train data on Big Data Platform such as Spark due to memory error on Python and Azure ML. Coincidently, it occurs memory error when loading Yelp dataset as well due to its large size. Transforming json file to csv file has shrunk the file size a lot. Also, dropping the review text column shrinks the file to less than 1GB. If there occurs memory error when training model, we will try dask package in python to deal with the large dataset. The paper has applied 3 classification models which are Logistic Regression, Decision Tree Regression and Gradient Boosting Tree Regression. The Gradient Boosting Tree Regression achieves the best result. In contrast, we will use decision tree and gradient boosting tree instead of regression version of them because our labels from dataset will be treated as 5 levels rather than continuous values.

Uniqueness

As mentioned earlier, there are lots of studies about Yelp review rating prediction. But most of them focused on sentiment analysis. In this project, it will not use any sentences from reviews, and it will predict results only by businesses' features and users' features. Previous studies can only predict results after customers have come to restaurants and written reviews; meanwhile, this project will predict review rating even when customers have not come to restaurant.

Dataset

The data sources will be observed from the Yelp dataset (Yelp Dataset 2019). As it provides JSON files with reliable data resources for the educational purposes. The observed data will be the backup of techniques and the tools using in the paper. And they will assist with the paper analysis. The Yelp dataset contains lots of restaurants and reviews. However, there are only 2 restaurants in Yelp dataset are located in Toronto, which is hard to analyze. As a result, this project would focus on customer reviews' rating on restaurant in British Columbia. There are 8635403 reviews which collected from 2004 to 2021. It means that there are lots of reviews written by 5 years ago, which may not have influence on current prediction, since many restaurants may be closed during 2004 to 2016. But we decide not to drop old review data, the reason is that the model will predict useless results on an existed restaurant if a customer's review on the same restaurant was written in 2010 but the review was deleted by data manipulation.

- Region: British Columbia, Canada
- The business.json, reviews .json and users.json will be the main datasets in this paper.
- The json files will be converted into csv files, and read as pandas dataframe.

Statistics of dataset

Business

	Business id	Business name	stars	Business in Vancouver, BC, CA	Business name in Vancouver, BC, CA
Number of unique values	160585	125850	9	10299	8946

User

	User id
Number of unique values	2189457

Review

	Review id	User id	Business id	Stars
Number of unique values	8635403	2189457	160585	5

	Date
Min	2004-10-12 11:14:43
Max	2021-01-28 15:38:54

Data description

Our dataset contains 3 tables which are business dataset, review dataset, and user dataset.

Business			Examples
business_id	String	22 character unique string business id	"tnhfDv5Il8EaGSXZGiuQGg"
name	String	the business's name	"Garaje"

Business			Examples
address	String	the full address of the business	"475 3rd St"
city	String	the city	"San Francisco"
state	String	2 character state code, if applicable	"CA"
postal_code	String	the postal code	"94107"
Latitude	float	latitude	37.7817529521
Longtitude	Float	longitude	-122.39612197
Stars	Float	star rating, rounded to half-stars	4.5
Review_count	Integer	number of reviews	1198
Is_open	Integer (1 or 2)	0 or 1 for closed or open, respectively	1

Business			Examples

Attributes	Dictionary	business attributes to values.
		note: some attribute values might be objects
			{
			"RestaurantsTakeOut": true,
			"BusinessParking": {
			"garage": false,
			"street": true,
			"validated": false,
			"lot": false,
			"valet": false
			},
			},

Categories	Array	an array of strings of business categories	["Mexican", "Burgers", "Gastropubs"]

Business			Examples

Hours	Dictionary	an object of key day to value hours, hours are using a 24hr clock	{
			"Monday": "10:00-21:00",
			"Tuesday": "10:00-21:00",
			"Friday": "10:00-21:00",
			"Wednesday": "10:00-21:00",
			"Thursday": "10:00-21:00",
			"Sunday": "11:00-18:00",
			"Saturday": "10:00-21:00"
			}
			}

Review			Examples
review_id	String	22 character unique review id	"zdSx_SD6obEhz9VrW9uAWA"
User_id	String	22 character unique user id, maps to the user in user.json	"Ha3iJu77CxlrFm-vQRs_8g"

Review			Examples
Business_id	String	22 character business id, maps to business in business.json	"tnhfDv5Il8EaGSXZGiuQGg"
Stars	Integer	star rating	4
Date	String	date formatted YYYY-MM-DD	"2016-03-09"
Text	Long-String	the review itself	"Great place to hang out after work: the prices are decent, and the ambience is fun. It's a bit loud, but very lively. The staff is friendly, and the food is good. They have a good selection of drinks."
Useful	Integer	number of useful votes received	0

Review			**Examples**
Funny	Integer	number of funny votes received	0
Cool	Integer	number of cool votes received	0

User			**Examples**
User_id	String	22 character unique user id, maps to the user in user.json	"Ha3iJu77CxlrFm-vQRs_8g"
Name	String	the user's first name	"Sebastien"
Review_count	Integer	the number of reviews they've written	56
Yelping_since	String	when the user joined Yelp, formatted like YYYY-MM-DD	"2011-01-01"

User			Examples
Friends	Array of strings	an array of the user's friend as user_ids	["wqoXYLWmpkEH0YvTmHBsJQ", "KUXLLiJGrjtSsapmxmpvTA", "6e9rJKQC3n0RSKyHLViL-Q"]
Useful	Integer	number of useful votes sent by the user	21
Funny	Integer	number of funny votes sent by the user	88
Cool	Integer	number of cool votes sent by the user	15
Fans	Integer	number of fans the user has	1032

User			Examples
Elite	array of integers	the years the user was elite	[2012, 2013]
Average_stars	Float	average rating of all reviews	4.31
Compliment_hot	Integer	number of hot compliments received by the user	339
Compliment_more	Integer	number of more compliments received by the user	668
Compliment_profile	Integer	number of profile compliments received by the user	42

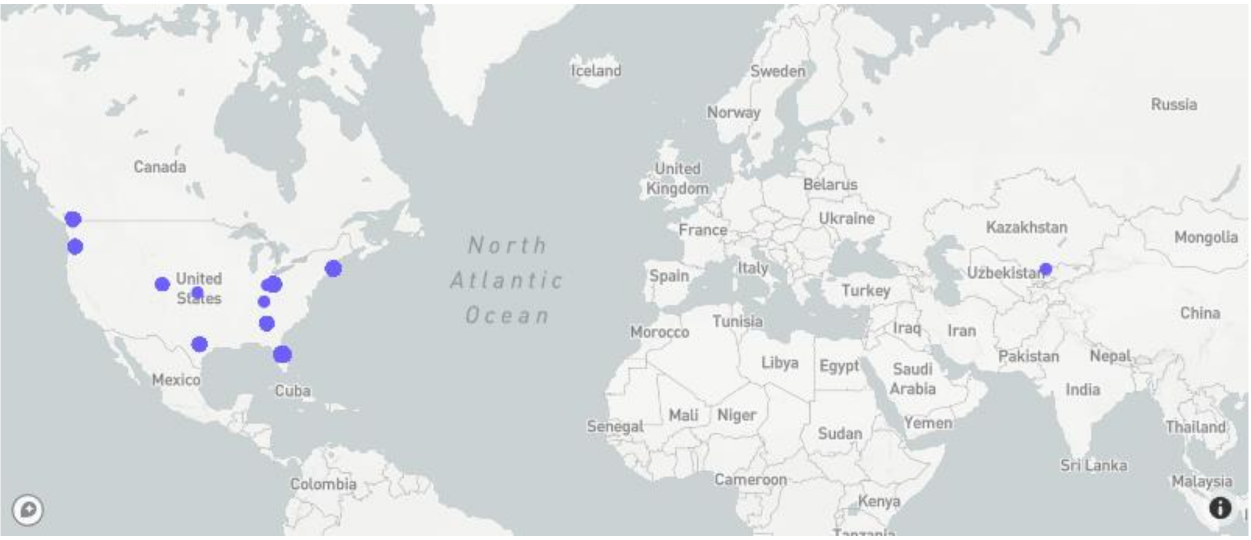
User			Examples
Compliment_cute	Integer	number of cute compliments received by the user	62
Compliment_list	Integer	number of list compliments received by the user	37
Compliment_note	Integer	number of note compliments received by the user	356
Compliment_plain	Integer	number of plain compliments received by the user	68
Compliment_cool	Integer	number of cool compliments received by the user	91
Compliment_funny	Integer	number of funny compliments received by the user	99

User			Examples
Compliment_writer	Integer	number of writer compliments received by the user	95
Compliment_photos	Integer	number of photo compliments received by the user	50

Initial Findings

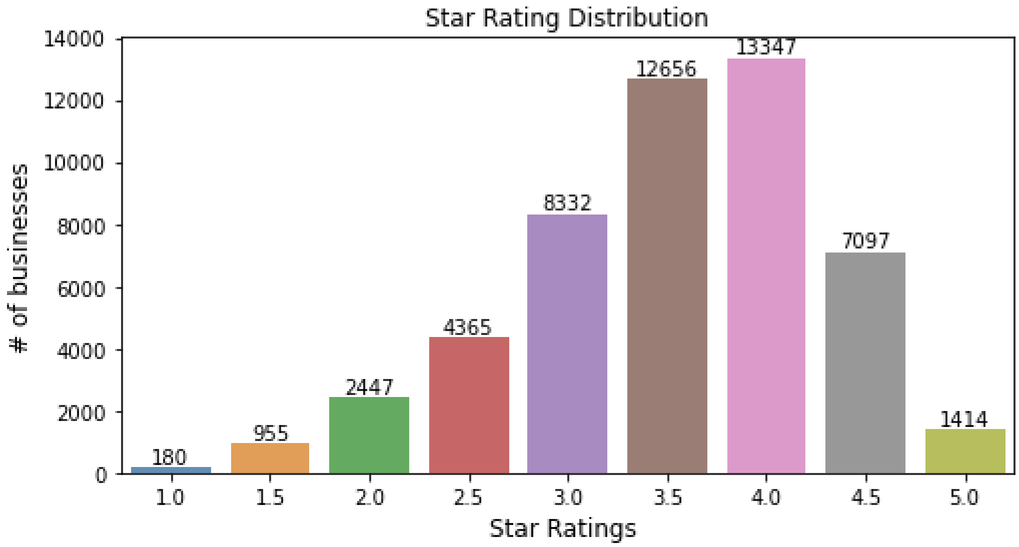
Map plot

In the map plot, restaurants were located by longitude and latitude. As shown in the map plot, the business data were concentrated on diverse big cities. Most restaurants mentioned in the dataset are located in the North America. Those centroid cities are Vancouver in Canada, Portland in the United States, Denver, Austin, Columbus, Atlanta, Orlando, Boston. It surprisingly finds that there are 2 Vancouver cities on where many restaurants are located, one was in Canada and the other one was in the United States. This project mainly focused on analyzing restaurants around the Great Vancouver Area in Canada.



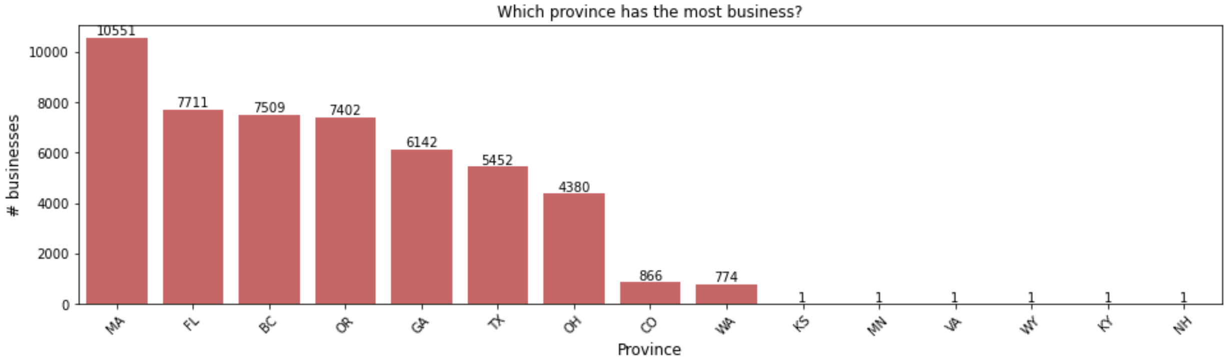
Rating star distribution

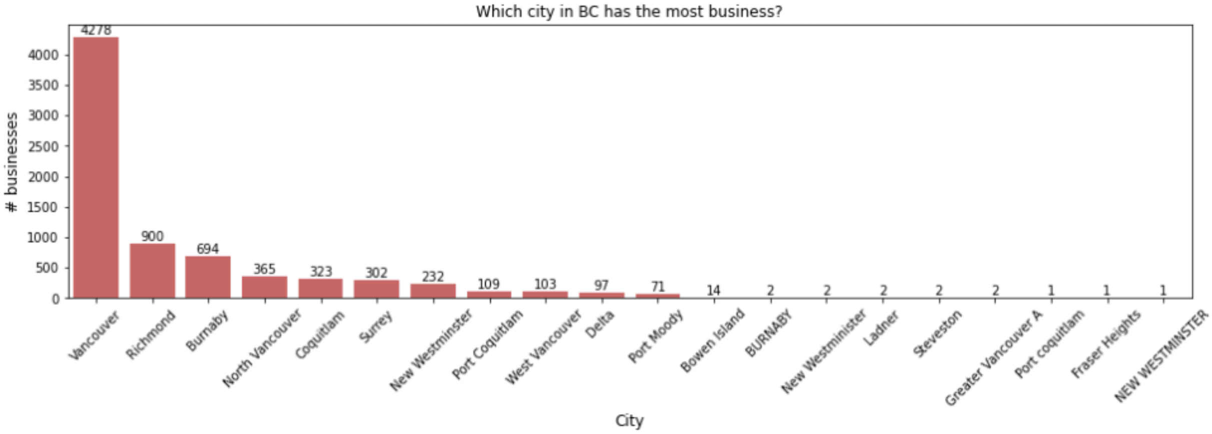
From the above histogram, it is not hard to see that the distribution of overall star ratings is skewed to the left. Most rating are lied between 3.0 to 4.5 stars, and 13347 restaurants are 4.0 stars which is the most common rating for restaurants. There are only 180 restaurants are 1 star and 1414 restaurants are 5 stars which commandeer only 3% of total restaurants. There is a common pattern on both 1-star and 5-star restaurants is that their number of reviews are pretty small, it may cause these restaurants' rating very high or very low rather than between the common range (3.0-4.5). And the review count for each restaurant will be discussed later.

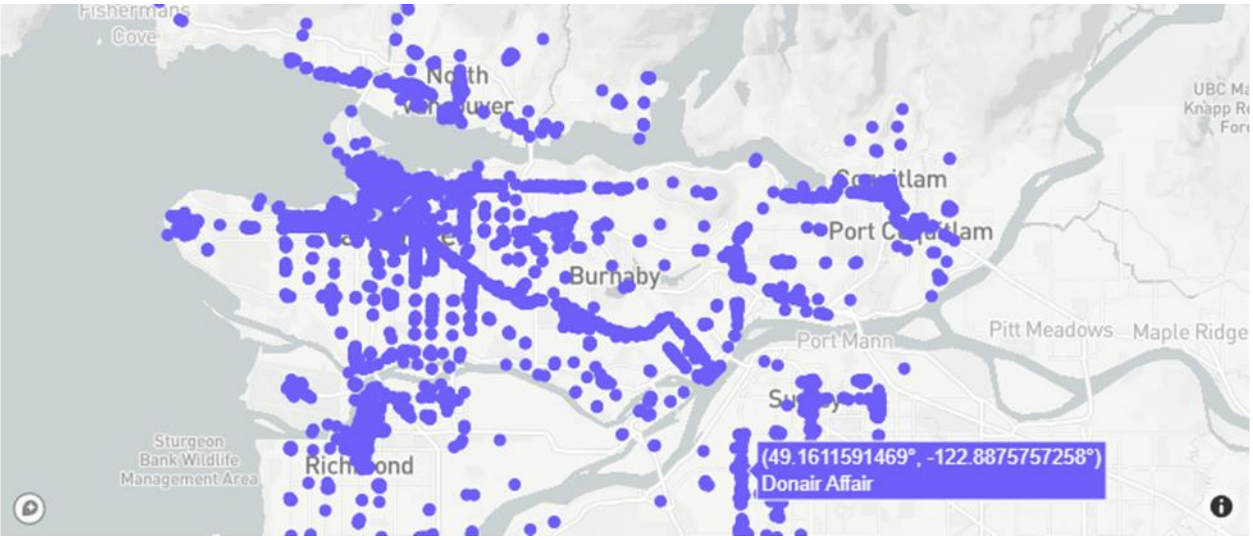


Business distribution

In the dataset, Massachusetts has around 10000 restaurants which is the most. Florida, British Columbia, Orlando followed as around 7000 restaurants. In the target area British Columbia, the main restaurants are located in Vancouver which is also can be discovered from map plot.







Top Restaurants in BC

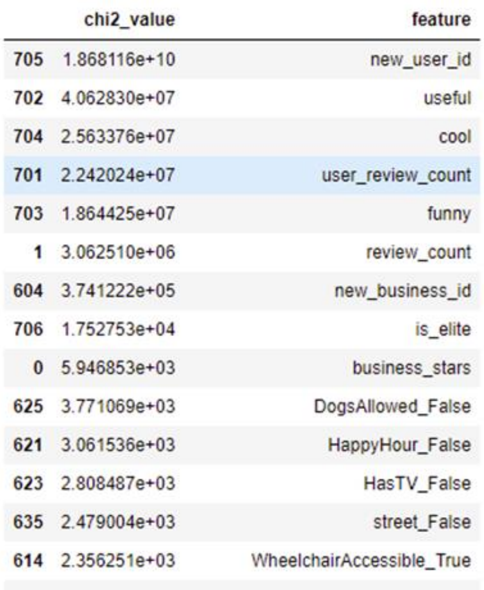
As mentioned earlier, most 5 stars restaurants have only a few reviews, which may cause biased rating. It is hard to determine the criteria for best restaurants, but it is not hard to find the top 10 popular restaurant. The top 10 popular restaurants are Medina Café, Miku, Chambar, Phnom Penh, Jam Café on Beatty, The Flying Pig- Yaletown, Joe Fortes Seafood & Chop House, Twisted Fork, Japadog, Hokkaido Ramen Santouka. These restaurants are all located in the Vancouver, and all their ratings are from 4 to 4.5 stars. These restaurants diverse from different themes, some of them are seafood restaurants while some of them provide brunch; however, the only common point between these restaurants is that they all have street parking. Most of them have bike parking and provide alcohol except Japadog, but Japadog is the only place allows dogs. After searching on internet, Japadog is a chained hot dog restaurant. In contrast, it is obvious that fine-dining restaurants always not allow dog to come in.

Features that most business have

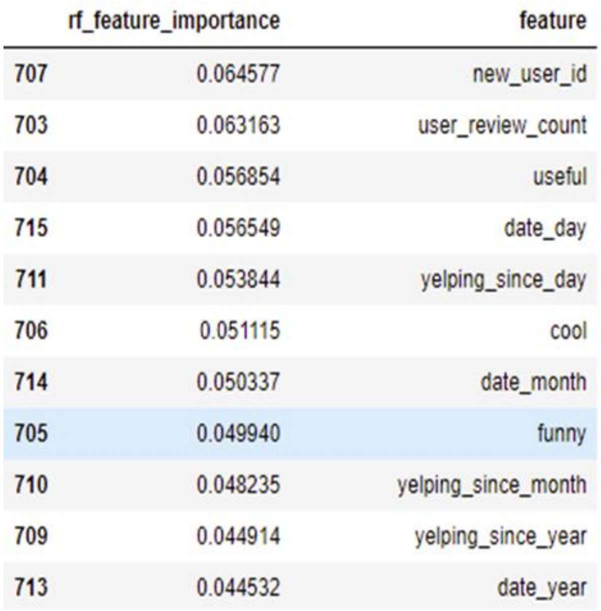
It is also need to find the features that most restaurants have, because it will figure out the foundation to run a restaurant. The dataset indicates that 95.07% restaurants have take-out service, 92.86% restaurants have wheelchairs to access, and 97.19% restaurants accept credit cards. These 3 criteria are fundamental for running business. There is an interesting founding when we compare ratio between British Columbia restaurants and all restaurants. 70.56% restaurants have table service and 52.14% restaurants have reservation service in British Columbia while only 58.52% restaurants have table service and 34.14% restaurants have reservation service in the world. It may suggest that table service and reservation service are important in British Columbia to run a restaurant.

Feature engineering

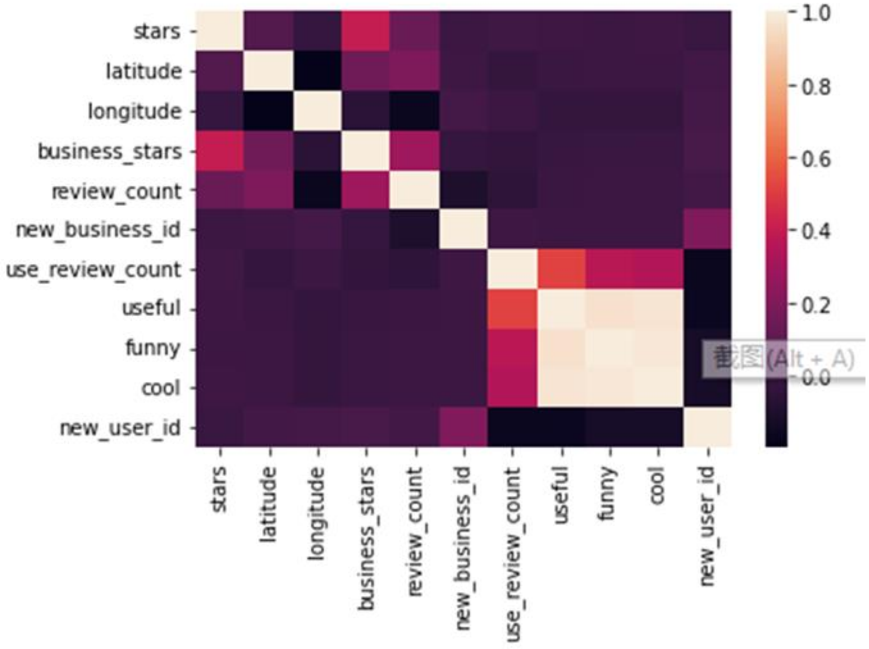
It is important to take a look of features. Since there are many characteristics for restaurants in the “categories” column, the final dataset has over 700 columns which is pretty large. The feature engineering process may help us to select features over 700 features, which not only reduce the model fitting time but also improve the result. First, the classical feature engineering method Variance Threshold has applied to the data, it computes the variance of each feature and eliminates the features whose variance is less than the threshold. If we set the threshold as 0.05, it will remain 108 columns. However, we decide not to use this method, because many of features are one-hot encoded. Considering most features in the dataset is categorical, it will apply Chi-square test and checking feature importance through random forest to the dataset. These 2 feature engineering methods are best for categorical data. The Chi-Square Test determine the relationship between 2 variables. It can be used to determine the relationship between features as independent variable and labels as dependent variable. The statistical formula for the Chi-Square test is the following: $x^2 = \sum(\frac{(O-E)^2}{E})$ (Pandis 2016). According to the table, user id, user review counts, business id, other users' opinion on the user's review (useful, cool, funny), users' Yelp membership is highly correlated to the review stars. The Chi-Square test can derive 50 to 100 features from original dataset, and the efficiency of filtered dataset will be discussed later. The Chi-Square test also have its limitation on value frequency. If the value frequency is pretty low in the column, the correlation may not be correct especially with one-hot encoded features. The Random Forest will also be applied to the dataset to check relationship between features and label. The information gain is a base concept in decision tree, which decides which feature should be used to split the data. The feature selection process through Random Forest built-in function feature importance is based on calculating information gain. According to the table, it is obvious that the highly correlated features are similar to the results from Chi-Square test. It is need to notice that the accuracy of feature importance from Random Forest model is relied on model itself, which means we can trust feature importance more if the Random Forest model is more accurate. The Random Forest model will also be discussed more in details later. At last, there is a heatmap plot shown the correlation between numeric features. From the plot, the rating label is related to business rating, which can be understood as that user is willing to give high rating to those high rating restaurants in common sense.



	chi2_value	feature
705	1.868116e+10	new_user_id
702	4.062830e+07	useful
704	2.563376e+07	cool
701	2.242024e+07	user_review_count
703	1.864425e+07	funny
1	3.062510e+06	review_count
604	3.741222e+05	new_business_id
706	1.752753e+04	is_elite
0	5.946853e+03	business_stars
625	3.771069e+03	DogsAllowed_False
621	3.061536e+03	HappyHour_False
623	2.808487e+03	HasTV_False
635	2.479004e+03	street_False
614	2.356251e+03	WheelchairAccessible_True



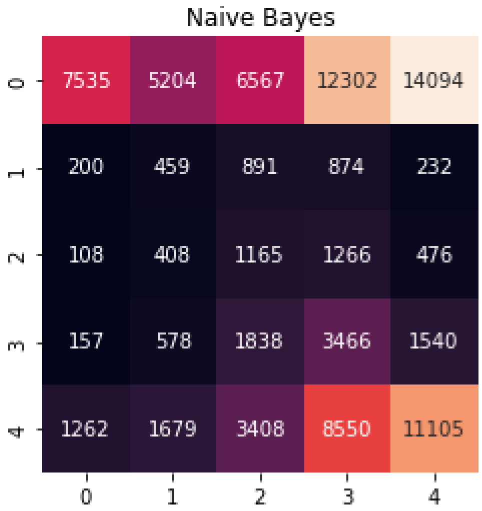
	rf_feature_importance	feature
707	0.064577	new_user_id
703	0.063163	user_review_count
704	0.056854	useful
715	0.056549	date_day
711	0.053844	yelping_since_day
706	0.051115	cool
714	0.050337	date_month
705	0.049940	funny
710	0.048235	yelping_since_month
709	0.044914	yelping_since_year
713	0.044532	date_year



Model results

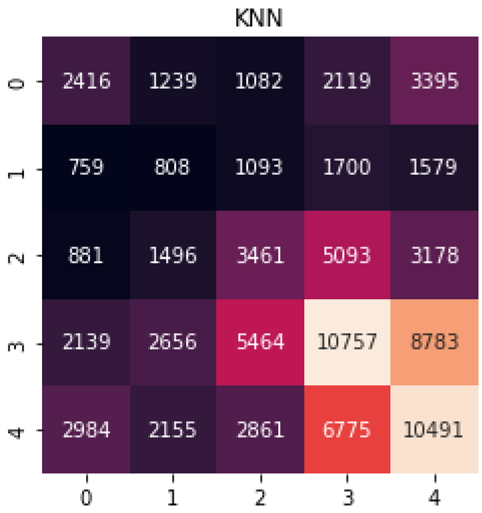
Naïve Bayes

Naive Bayes is one of the popular machine learning computer algorithms that are able to apply on a broad range of classification jobs. Normal implementations include filtering spam, document classification, sentiment prediction, etc. It is helpful with the simple structure and useful for large datasets. Moreover, Naive Bayes is well known for surpassing extremely complex classification procedures. It is a classification method that resource from the Bayes theorem, there is a significant hypothesis about the indicator’s independence. It uses category-specific statistics to calculate the ratio between events. Bayes' theorem determines the subsequent probability. In other words, an occurrence of a specific feature in a category not related to the existence of another feature. In real-world data concerns, we usually have several X variables. The term ‘Naive’ is reasonable since it has the naive hypothesis that the X’s must be independent of all else. It can simply and quickly predict the category of data batches. Also, it can forecast several different categories at the same time when using multinominal objective function. Naive Bayes does it more efficiently than other models. For example, other models like logistic regression require more training data. After the one-hot encoding, all categorical data has transferred to numerical columns. Since the label was composed of 5 classes, it requires Multinominal Naïve Bayes for modelling. When we fitted the model with our original data, the accuracy of Naïve Bayes model is pretty low around 16%. And there does not exist any predictions on 2 stars. It needs to notice that Business ID and User ID are high correlated to other features in original dataset, but the Naïve Bayes model assumes that all features are independent to each other because the Bayes Theorem can only be applied to two events that are independent. As a result, we decide to drop the business_id and user_id features, and fit the model again. After dropped two columns, the accuracy of model has improved to 27.40% which leaped a lot from original result. Since the Naïve Bayes cannot handle continuous values, it will help model to predict if dropping the longitude and latitude columns. After dropped two continuous columns, the accuracy increased to 27.56%, and it takes 11.34 seconds to fit the model in total. Then, fit the model with 50 features selected by random forest without business_id, user_id, longitude and latitude, the accuracy increased to 27.80%; meanwhile, the fitting time decreased to 0.8 seconds which is pretty fast. Fitted the Naïve Bayes with selected features cost less time but achieve a better result, which implies that the feature engineering is successful at this moment. The confusion matrix of final result is attached below.



KNN

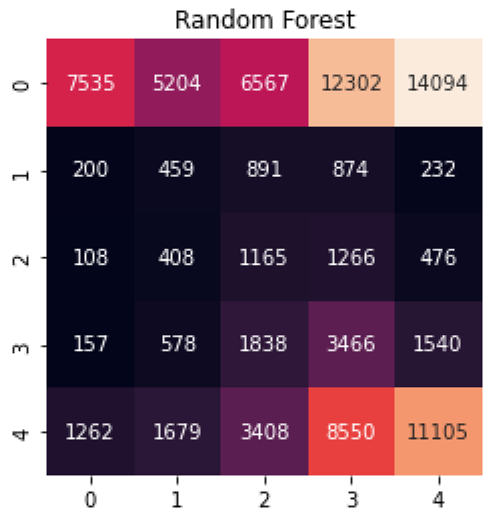
KNN can be utilized for primarily categorization problems. It also called K nearest neighbor, which is a simple computer algorithm that keeps all accessible cases and categorized the volume of new works by a most vote of its k neighbours. The case is assigned to the category that cooperates best among its K nearest neighbors, and is calculated by the distance function. Various distance functions are applicable, such as Euclidean, Manhattan, Minkowski and Hamming. Euclidean, Manhattan, and Minkowski functions are employed for the uninterrupted task while Hamming function works for classification variables. If K = 1, then the case is only assigned to the category of its close neighborhood. Hence, the selection of K becomes challenging by using KNN. In this project, K was chosen as 5 to check initial result. The KNN is an expensive model as a lazy learner, it costs a lot of time to calculate the distance between each point in dataset, Since the dataset contains over 400000 rows, the prediction time for KNN is huge. When using original dataset to predict the result, it cost about one hour to get the result when choosing K as 5. As expected, the accuracy for KNN model is pretty good, which achieves 30.22%. The problem with KNN is that it costs too much time, and the problem may be solved if fitting the model with selected features. When fitted the model with 50 features selected through Random Forest without business_id and user_id, the accuracy has improved to 32.72%, and the prediction time reduced to 10 minutes. It suggested that the feature selection can also help KNN as well.



	0	1	2	3	4
0	2416	1239	1082	2119	3395
1	759	808	1093	1700	1579
2	881	1496	3461	5093	3178
3	2139	2656	5464	10757	8783
4	2984	2155	2861	6775	10491

Random forest

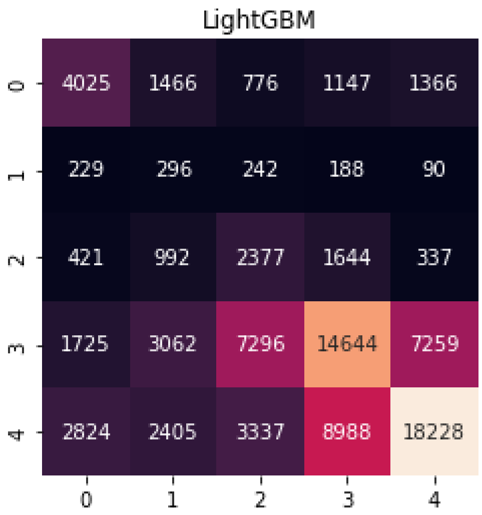
Random Forest operates for the purpose of collecting decision trees. At Random Forest, we are gathering decision trees. To categorize a new target by different features, the respective tree allows a categorization and we declare the tree "votes" for this category. The forest selects the category with the greatest number of votes. Respective tree is grown to the broader scale without trimming. The advantages of Random Forest are various. Which includes ease of applying, productivity, veracity and beginner welcoming. This is why we use Random Forest to select features as mentioned earlier in this report. In this part, we will focus on how it performs on predicting results. Since the Random Forest will select feature by itself, we did not need to eliminate features before fitting the model. However, in order to save the fitting time at first, the 50 selected features have been used to fit the model. When choosing number of estimators as 1000, the Random Forest achieves a high accuracy at 35.52% and takes 15 minutes to fit the model. Since the accuracy is good comparing to other models, it worth time to fit the Random Forest with all features. When the model fitted with whole features, the modeling process takes 45 minutes, but the accuracy becomes 34.09% this time. It implies that it is better to fit model with less features according to feature importance selection to avoid overfitting. In addition, it should be noticed that Random Forest can handle overfitting well by the concept of bagging.



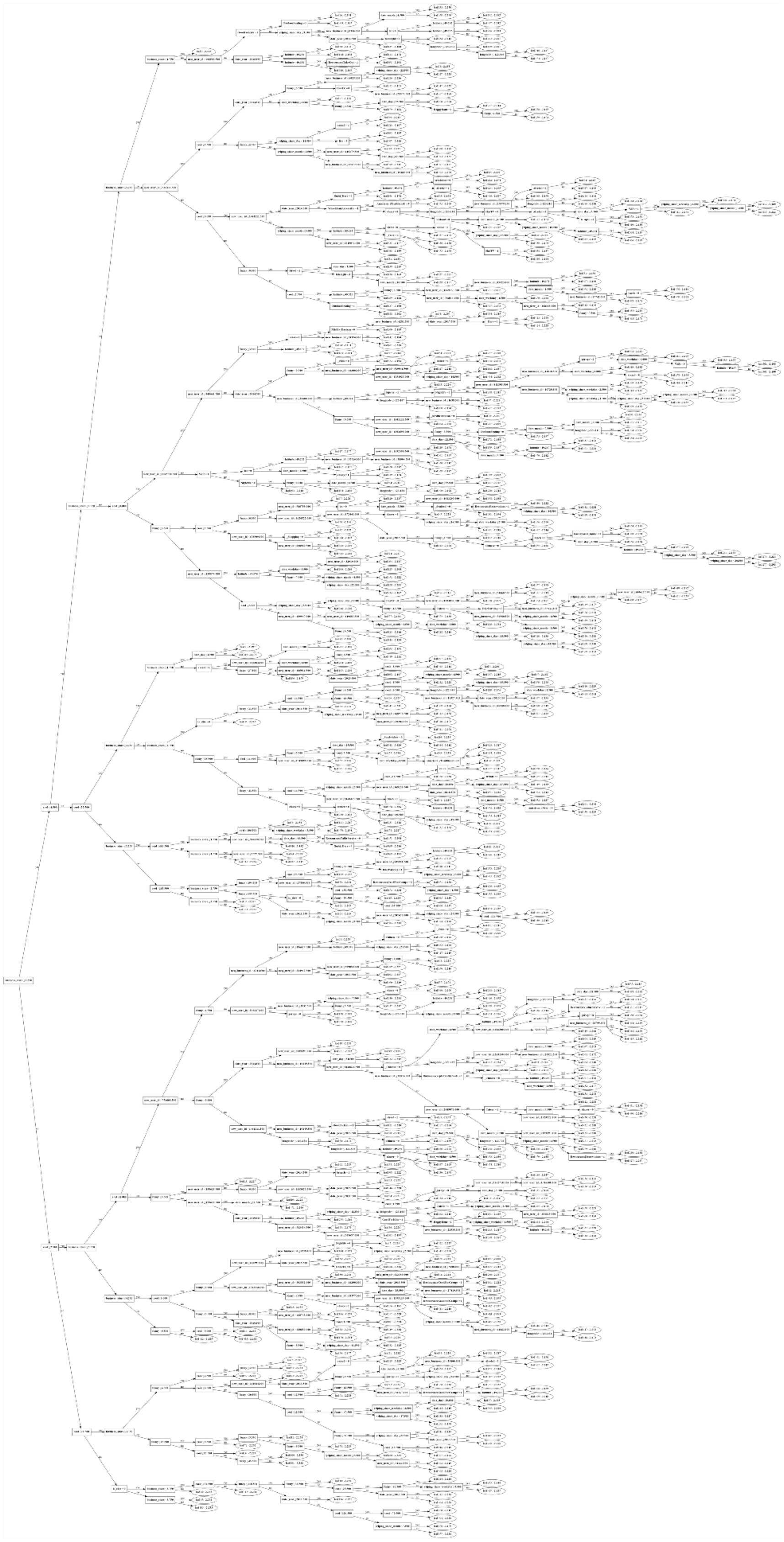
	0	1	2	3	4
0	7535	5204	6567	12302	14094
1	200	459	891	874	232
2	108	408	1165	1266	476
3	157	578	1838	3466	1540
4	1262	1679	3408	8550	11105

LightGBM

LightGBM is an increasingly inclined framework based on tree-based learning classifiers. The structure is quick and efficient for classifying, categorizing, and many machine learning tasks. LightGBM classifies the leaves with the most seizures as wise, while other classifiers classify the size or level of the tree as wise rather than leaf wise. Therefore, when the same leaf is developed in LightGBM, compared with other classifiers, leaf-based computer algorithms can further reduce waste. Comparing to other GBDT models like XGBoost, LightGBM has better efficiency and scalability with high dimension and large dataset. Its benefits are also reflected in accelerating training speed and increasing productivity, lowing memory usage, parallel and GPU learning backup, and efficient processing of integrated data. Since LightGBM can handle categorical data, and achieve a better result with categorical value rather than one-hot encoding the training data for LightGBM have not applied one-hot encoding. The biggest problem for Random Forest is that it is too expensive to fit in terms of time value. The LightGBM uses leaf-wise learning rather than level-wise learning, which can reduce a lot of time for splitting unnecessary nodes. Gradient-based One-Side Sampling (GOSS) makes the model only calculate information gain on those features which has large gradient. As a result, the modeling time for LightGBM has reduced a lot. As noticed that, the original dataset has 665 columns; fortunately, Exclusive Feature Bundling can partition features into small bundles to reduce dimensionality. Theoretically, the LightGBM will be excellent for the question, which will be proved by actual results in the next. The loss function for this data is Soft-Max function, which can be simply interpreted as : $\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$. With only 500 rounds, the LightGBM has achieved 36.92% accuracy in only 6minutes and 20 seconds. It to be noticed that, this model has only 54.11 accuracy on training set, which means it have not been overfitting and it has a lot of spaces to learn. With 5000 rounds, it shows that the loss on training set keeps going down, but the loss on validation set goes down in first 1000 iterations then go up till the end. The accuracy of testing set is 37.09% while the training set's accuracy arrives 99.17%. In order to get rid of overfitting, we reduced the number of rounds to 900, and the accuracy on testing set becomes 37.17% when we limit the number of leaves at 600. And the fitting time has been improved from 45 minutes to 9 minutes. We have tried to tune hyperparameters several times, but the accuracy remains around 37.1%, and there does not exist a distinct difference. As a result, the best LightGBM model will stay as a multiclass objective function has learning rate at 0.01, maximum leaves at 500, feature fraction at 0.9, number of rounds at 1000 model, with 37.20% accuracy. There is no big difference on feature importance between Random Forest and LightGBM. For both methods, features like “review count”, “user review count”, “WiFi”, “Restaurant delivery”, “Noise Level” are important features. Here is one plot of 4500 decision trees.



	0	1	2	3	4
0	4025	1466	776	1147	1366
1	229	296	242	188	90
2	421	992	2377	1644	337
3	1725	3062	7296	14644	7259
4	2824	2405	3337	8988	18228



Conclusion

In this paper, we have discovered the basic business information behind Yelp dataset, and compared 4 classification models: Naïve Bayes, KNN, Random Forest, and LightGBM. We have performed both theoretical and experimental studies on these 4 models. The experimental results are consistent with expectations how these 4 models perform. Comparing the final results of these 4 models, the LightGBM wins by its accuracy and high efficiency. The accuracy of model has achieved 37.17% which is the best, and it only take 9.5 minutes which is fast comparing to Random Forest. After checking the feature importance, the user's review count and business's review count and the number of useful reviews are high correlated to review rating stars. In terms of running business, alcohol, restaurant price range, WiFi, delivery, outdooring seating are important for users to rate a restaurant.

Future Plan

The Yelp dataset is really large, and there is huge information behind the dataset. This report mainly focused on British Columbia restaurants, which is a small part of dataset. In the future, the analyzing region can be extended to whole world. With more data, it should take more time to fit model and achieve more accurate results. In terms of restaurants, there also exist a cold start problem. For those restaurants and users with few reviews, the model cannot predict the rating based on missing past data. There should exist another model to deal with cold start problem. In our current dataset, there are 58226 users who have only 1 review on BC restaurants, which accounts for 54.59% of all unique users. All BC restaurants have at least 5 reviews, but it is not enough to judge a restaurant. So, the cold start problems have occurred both on restaurants and users. For model fitting, although LightGBM achieves a good result, we believe Deep Learning model may improve accuracy as well. Several neural network models will be tried in the future. Even for LightGBM, there are so many features to tune, which may improve the result a little bit, but will not have a huge leap.

Reference

1. Yelp Dataset. (2019). Yelp.com. https://www.yelp.com/dataset
2. E. S. Alamoudi and S. A. Azwari. (2021). "Exploratory Data Analysis and Data Mining on Yelp Restaurant Review," 2021 National Computing Colleges Conference (NCCC), pp. 1-6, doi: 10.1109/NCCC49330.2021.9428850.
3. Nakayama, M., & Wan, Y. (2018, September 6). *The cultural impact on social commerce: A sentiment analysis on yelp ethnic restaurant reviews*. Information & Management. Retrieved October 19, 2021, from https://www.sciencedirect.com/science/article/abs/pii/S0378720617306225.
4. Luo, Y., & Xu, X. (2019). Predicting the Helpfulness of Online Restaurant Reviews Using Different Machine Learning Algorithms: A Case Study of Yelp. *Sustainability*, *11*(19), 5254. http://dx.doi.org/10.3390/su11195254
5. Ariyasriwatana, W., Buente, W., Oshiro, M., & Streveler, D. (2014). Categorizing health-related cues to action: using Yelp reviews of restaurants in Hawaii. New Review of Hypermedia and Multimedia, 20(4), 317–340. doi:10.1080/13614568.2014.987326
6. Q. Xuan, M. Zhou, Z. Zhang, C. Fu, Y. Xiang, Z. Wu, & V. Filkov. (2018). Modern Food Foraging Patterns: Geography and Cuisine Choices of Restaurant Patrons on Yelp. IEEE Transactions on Computational Social Systems, 5(2), 508–517. doi:10.1109/TCSS.2018.2819659
7. Keller, D., & Kostromitina, M. (2020). Characterizing non-chain Restaurants' Yelp Star-Ratings: Generalizable findings from a representative sample of yelp reviews.

International Journal of Hospitality Management, *86*, 102440. https://doi.org/10.1016/j.ijhm.2019.102440

8. Liu, Z. (2020, December 12). *Yelp review rating prediction: Machine learning and deep learning models*. arXiv.org. Retrieved October 19, 2021, from https://arxiv.org/abs/2012.06690.
9. A. Rafay, M. Suleman and A. Alim. (2020). "Robust Review Rating Prediction Model based on Machine and Deep Learning: Yelp Dataset," 2020 International Conference on Emerging Trends in Smart Technologies (ICETST), pp. 8138-8143, doi: 10.1109/ICETST49965.2020.9080713.
10. Wu, S., Zhang, Y., Zhang, W., Bian, K., & Cui, B. (2021). Enhanced Review-based rating prediction by exploiting aside information and user influence. *Knowledge-Based Systems*, *222*, 107015. https://doi.org/10.1016/j.knosys.2021.107015
11. Wang, B., Chen, B., Li, M., & Zhou, G. (2019). User-Personalized Review Rating Prediction Method Based on Review Text Content and User-Item Rating Matrix. *Information*, *10*(1), 1. http://dx.doi.org/10.3390/info10010001
12. Woo, J., & Mishra, M. (2020). Predicting the ratings of Amazon products using Big Data. *WIREs Data Mining and Knowledge Discovery*, *11*(3). https://doi.org/10.1002/widm.1400
13. Pandis, N. (2016). The chi-square test. *American journal of orthodontics and dentofacial orthopedics*, *150*(5), 898–899. doi:10.1016/j.ajodo.2016.08.009