YELP REVIEWS

Use of ML and NLP to analyze customer satisfaction

“submitted towards partial fulfilment of the criteria for award of PGP-DSE by GLIM”

SUBMITTED BY

STUDENT NAME	ROLL NUMBER
Brijesh	DSEFTBNOV18004
Debashis Gogoi	DSEFTBNOV18019
S T Mohammed	DSEFTBNOV18024
Sivavamsi	DSEFTBNOV18014

BATCH: DSE_BLR_NOV2019

MENTOR: Mrs. C.R.SOWMYA

GREAT LAKES
INSTITUTE OF MANAGEMENT
Global Mindset - Indian Roots

ABSTRACT AND KEYWORDS

ABSTRACT:

The project involves identifying the unhappy customers and also the factors which lead to their dissatisfaction. We create various supervised models to identify the model with the highest Specificity value to reduce the effect of causing a Type II error so that we don’t lose our valuable customers.

We create supervised models on 2 types of data, one is using our structured data using the attributes given in the structured form and the 2nd is using our reviewer reviews and performing text classification and building models based on text.

The prediction is conducted by using machine learning techniques like Logistic Regression, Decision Tree Classifier, Random Forest Classifier, GaussianNB and MultinomialNB. The project provides insights about the significant factors affecting the businesses.

KEYWORDS:

Customer Satisfaction Level Prediction, Machine Learning – Supervised Models, Text Classification, NLP, Text Mining, Logistic Regression, Decision Tree, Random Forest, GaussianNB, MultinomialNB.

ACKNOWLEDGEMENTS

At the outset, we are indebted to our Mentor Mrs. Sowmya for her time, valuable inputs and guidance. Her experience, support and structured thought process guided us to be on the right track towards completion of this project.

We are grateful to the course directors – DSE Program. Their in-depth knowledge coupled with their passion in delivering the subjects in a lucid manner has helped us a lot. We are thankful to them for the various sessions they have provided.

We are thankful to Mrs. Uma, Program Manager DSE Program for her unflinching and unabated help extended to us always.

We also thank all the course faculty of the DSE program for providing us a strong foundation in various concepts of analytics and machine learning.

We would like to thank our respective families and friends for giving us the necessary support to complete this program.

CERTIFICATION OF COMPLETION

I hereby certify that the project titled “YELP REVIEWS - Use of ML & NLP to analyse customer satisfaction” was undertaken and completed under my guidance and supervision by Brijesh, Debashis, Mohammed and Vamsi, students of the November 2019 batch of the Post Graduate Program in Data Science Engineering, Bangalore.

Mrs. C.R.Sowmya
Date: 3rd April 2019

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Background and need for study:

Yelp is an American multinational corporation founded in 2004 by two former PayPal employees headquartered in San Francisco, California. It develops, hosts and markets Yelp.com and the Yelp mobile app, which publish crowd-sourced reviews about local businesses, as well as the online reservation service Yelp Reservations. Yelp grew quickly and raised several rounds of funding. By 2010 it had $30 million in revenues and the website had published more than 4.5 million crowd-sourced reviews. Yelp became a public company in March 2012 and became profitable for the first time two years later in 2014. Yelp has 77 million unique visitors via desktop computer and 64 million unique visitors via mobile website on a monthly average basis. By the April, 2018, Yelp had 171 million reviews.

Scope & Objectives:

Our aim is to flag dissatisfied customers based on review text, review rating and quality of reviewer so that any dissatisfied customer is immediately brought to the attention of the business for them to take relevant action. Apart from this we intend to perform sentiment analysis based on the reviewer text to get an insight of where a particular review falls in.

Approach & methodology:

This project uses the approach of text mining on the review text to classify reviews based on satisfied and dissatisfied customers. The text mining also helps in identifying aspects that make a customer happy or unhappy.

CHAPTER-1 PROJECT OVERVIEW

Yelp is an American multinational corporation which develops and hosts website and mobile app that publish crowd sourced reviews about local businesses and online reservation service. With over 5 million reviews available publicly, Yelp is a rich data source to understand which features of a business contribute more towards its success. A business is successful if it has a good reputation and is able to generate profit. Reputation of a business is the reflection of customers' opinion based on their previous experience or based on a common image of a particular business.

Problem Objective

Objective is to identify the dissatisfied customers using text mining.

Description

Yelp is an American multinational corporation founded in 2004 by two former PayPal employees headquartered in San Francisco, California. At present Yelp has around 77 million unique visitors via desktop computer and 64 million unique visitors via mobile website on a monthly average basis. By the April, 2018, Yelp had 171 million reviews.

Problem Statement

Objective is to flag dissatisfied customers based on review text, review rating and quality of reviewer so that any dissatisfied customer is immediately brought to the attention of the business for them to take relevant action.

Domain

Retail

Data Source

https://data.world/brianray/yelp-reviews/workspace

Data Information

Yelp_training_set_review.csv:

- The dataset contains of 229907 rows and 31 columns
- This dataset contains all the useful attribute needed for analysis.
- yelp_training_set_busines.csv and yelp_training_set_user.csv both datasets are merged into this dataset.

The attribute names and a little description about each of them are:

No	Name of attribute	Description
1	business_blank	this attribute contains only one value of boolean type 'FASLE'
2	business_categories	a local body listed on yelp like Restaurants, Department Stores, Bars, Home -Local Services, Cafes, Automotive, etc. which are expressed as what the type of business it is attached to.
3	business_city	the names of city based in the business is based
4	business_full_address	address of the business
5	business_id	a business unique 'ID' to uniquely identify the business
6	business_latitude	latitude of where a business is located
7	business_longitude	longitude of where a business is located
8	business_name	name of the business
9	business_neighborhoods	contains 100% missing data
10	business_open	containing 2 boolean values as 'TRUE' or 'FALSE', this attribute refers to whether a business is running or it is closed
11	business_review_count	it the number of reviews a business has received
12	business_stars	rating of a business (unknown source from how this rating is given)
13	business_state	state in the US where a business is located
14	business_type	with only 'business' as a value in all rows
15	cool	contains values from 0 to 117
16	date	date on which a review is given
17	funny	contains values from 0 to 70
18	review_id	unique id given to each review
19	reviewer_average_stars	average stars of each reviewer
20	reviewer_blank	contains 2 boolean values as 'TRUE' or 'FALSE'

21	reviewer_cool	this is a rating given by other users to a review to express how cool his review is
22	reviewer_funny	this is a rating given by other users to a review to express how funny a review is
23	reviewer_name	name of a reviewer
24	reviewer_review_count	count of reviews given by a reviewer
25	reviewer_type	contains ‘user’ as the only value across all rows
26	reviewer_useful	this is a rating given by other users to a review to express how useful a review is
27	stars (target attribute)	ratings varying from a minimum of 1 to a maximum of 5 as given by users to a business
28	text	It is text written by user after visiting business about the overall experience. It also gives a numeric representation (out of 5) to compare it with other business which is named as ‘stars’.
29	type	contains ‘review as the only value across all rows
30	useful	contains values from 0 to 120
31	user_Id	A person’s unique ‘ID’ who has registered on yelp who is writing reviews about different business after visiting them or a person who is using yelp reviews to choose business.

Data Cleaning:

Data cleansing or data cleaning is the process of detecting and correcting (or removing) missing, corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.

Missing Data Treatment:

- With our data of almost 250 mb, we have 229907 rows and 31 columns. The first step we did in data cleaning is checking for the presence of missing values if any. Our result gave us 4 columns containing missing data.
- “**business_neighborhoods**”attribute has 100% missing data, so the whole of the column is dropped
- Coming down to the attribute “**business_categories**” which had 777 is only 0.33% of our overall data, so we drop those 777 rows.

- The “**text**” column containing 6 rows of missing values are also dropped.
- The final attribute with null values “**reviewer_name**” is also dropped with the view that it is already represented by the attribute “user_id” which uniquely identifies a reviewer.

Data Cleaning other than missing value treatment:

- **business_blank** column contained only one value of boolean type “FALSE” so we drop this column
- **reviewer_type** contains ‘user’as the only value for all the records so we dropped the column
- **type** column contains‘review’ as the only value for all therecords so we dropped the column
- **reviewer_name** column is dropped because we have user_id column both indicating same meaning
- **business_name**is dropped because we will use ‘business_id’ as the alterate column against a business name which will uniquely define a business
- We also drop few other columns such as **cool**, **funny**, **useful**, **reviewer_type** as these neither infer any meaning nor help us in any way for our analysis.
- Another major data cleaning part we did is in **business_state**, we had 4 states, namely AZ, CA, CO and SC. Wherein CA, CO and SC contributes to just 16 records in total. So we remove the records having these countries in **business_state** and will focus only on the records in the state of AZ.

After all of our data cleaning process, we were finally left with 215152 rows and 21 columns.

Final list of attributes after all of the data cleaning process are:

business_categories	business_city	business_full_address
business_id	business_latitude	business_longitude
business_open	business_review_count	business_stars
business_state	date	review_id
reviewer_average_stars	reviewer_blank	reviewer_cool
reviewer_funny	reviewer_review_count	reviewer_useful
stars	text	user_id

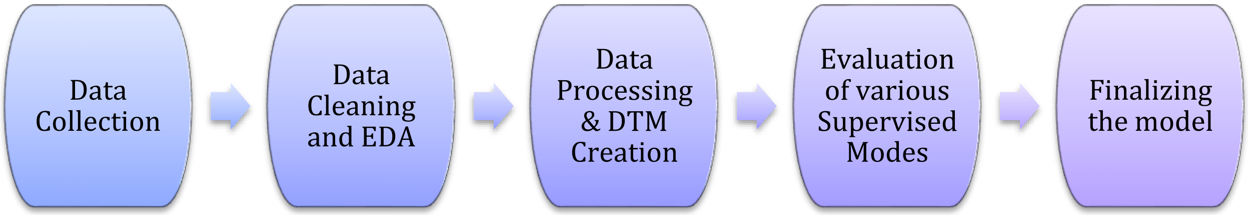
Feature Engineering

Feature engineering is the process of using **domain knowledge** of the data to create features that make machine learning algorithms work. If feature engineering is done correctly, it increases the predictive power of machine learning algorithms by creating features from raw data that help facilitate the machine learning process.

In our Yelp data, we create a number of new features so as to understand the data more other than just visualizing. Here we talk about them :

1. We created a new attribute, where we segregated all the business categories to 16 unique type of business by using certain text which were present in the business categories.
2. The 2nd new attribute we added is to understand the quality of a reviewer as stated in our business problem. We did this by creating 5 bins, based on the number of reviews a reviewer has given and assigned 5 different scores to each bin like as score of 1 if a reviewer has a review count between 0-100, a score of 2 in a range of 101-200 and so on an the maximum score being 5 for reviews more than 400.
3. The 3rd and the 4th new attribute added is the number of words in each review and the number of characters in each review respectively. We this to check if the length of a review contributes or helps us to understand of how ratings are being given.

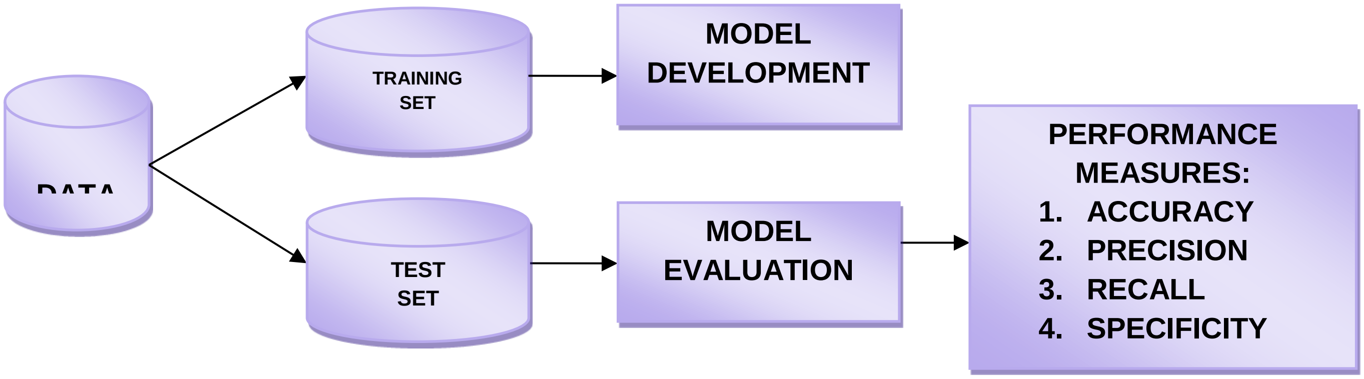
Approach Followed


Data Collection
Data Cleaning and EDA
Data Processing & DTM Creation
Evaluation of various Supervised Modes
Finalizing the model

Workflow for a classification problem

Whenever we perform classification, the first step is to understand the problem and identify potential features and label. Features are those characteristics or attributes which affect the results of the label. These characteristics are known as features which help the model classify customers.

The classification has two phases, a learning phase, and the evaluation phase. In the learning phase, classifier trains its model on a given dataset and in the evaluation phase, it tests the classifier performance. Performance is evaluated on the basis of various parameters such as accuracy, precision, recall and specificity



Statistical Tools and Techniques

In machine learning and statistics, classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known. In the terminology of machine learning, classification is considered an instance of supervised learning, i.e., learning where a training set of correctly identified observations is available.

An algorithm that implements classification, especially in a concrete implementation, is known as a classifier. The term "classifier" sometimes also refers to the mathematical function, implemented by a classification algorithm, that maps input data to a category.

In statistics, where classification is often done with logistic regression or a similar procedure, the properties of observations are termed explanatory variables (or independent variables, regressors, etc.), and the categories to be predicted are known as outcomes, which are considered to be possible values of the dependent variable. In machine learning, the observations are often known as *instances*, the explanatory variables are termed *features* (grouped into a feature vector), and the possible categories to be predicted are *classes*.

Classification can be thought of as two separate problems – binary classification and multiclass classification. In binary classification, a better understood task, only two classes are involved, whereas multiclass classification involves assigning an object to one of several classes. Since many classification methods have been developed specifically for binary classification, multiclass classification often requires the combined use of multiple binary classifiers.

Here, in our project, we treat the attribute ‘stars’ as classes or dependent variable which contains values for five (5) different classes, namely 1, 2, 3, 4 and 5. Which implies the rating of a business as given by a user. 1 meaning the least rating a user can give and 5 being the best and the highest. Our aim here is to correctly predict the rating or star a user will give to a business based on how one feels. So herein, our independent variable or instance will be the attribute ‘text’ or user review. This attribute will contain words based on which we will build a model to predict the class of star where a business will fall or how much star/rating a business can attain based on the review as given by a user.

Our problem here is a multiclass classification problem as we have 5 different classes of ratings and we intend to predict only for star ratings 1 and 5.

There are many classification models available, here we will be talking about a few of them which we plan to use in our project for better evaluation of our model. Here are the classification models we shall use here for our analysis:

1. Logistic Regression
2. Decision Tree
3. Random Forest
4. Gaussian Naïve Bayes
5. Multinomial Naïve Bayes

Performance Metrics for Classification problems

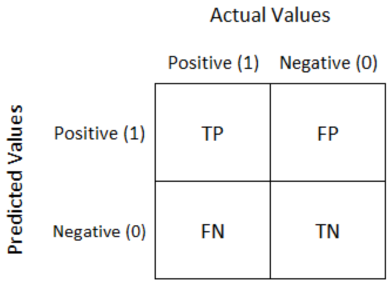
The metrics that we choose to evaluate our machine learning model is very important. Choice of metrics influences how the performance of machine learning algorithms is measured and compared. Here in our problem we will discuss about the various classification model performance measures.

1. Confusion Matrix
2. Accuracy
3. Precision
4. Recall or Sensitivity
5. Specificity

Let us now discuss about these performance measures for a classification problem:

Confusion Matrix:

The Confusion matrix is one of the most intuitive and metrics used for finding the correctness and accuracy of the model. It is used for Classification problem where the output can be of two or more types of classes.



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

The Confusion matrix in itself is not a performance measure as such, but almost all of the performance metrics are based on Confusion Matrix and the numbers inside it.

True Positives (TP): True positives are the cases when the actual class of the data point was 1(True) and the predicted is also 1(True).

True Negatives (TN): True negatives are the cases when the actual class of the data point was 0(False) and the predicted is also 0(False).

False Positives (FP): False positives are the cases when the actual class of the data point was 0(False) and the predicted is 1(True). False is because the model has predicted incorrectly and positive because the class predicted was a positive one. (1)

False Negatives (FN): False negatives are the cases when the actual class of the data point was 1(True) and the predicted is 0(False). False is because the model has predicted incorrectly and negative because the class predicted was a negative one. (0)

Accuracy:

Accuracy in classification problems is the number of correct predictions made by the model over all kind’s predictions made. Accuracy is a *good* measure when the target variable classes in the data are nearly balanced. Accuracy should *never* be used as a measure when the target variable classes in the data are a majority of one class.

$$\textbf{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision:

In simple terms, precision can be defined as how many selected items are relevant. Precision is about being precise. So even if we managed to capture only one case, and we captured it correctly, then we are 100% precise.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall or Sensitivity:

Recall or Sensitivity in simple terms can be defined as how many relevant items are selected. Recall is not so much about capturing cases correctly but more about capturing all cases with the answer.

$$\text{Recall} = \frac{TP}{TP + FN}$$

Specificity:

Specificity (also called the true negative rate) measures the proportion of actual negatives that are correctly identified as such.

$$\text{Specificity} = \frac{TN}{TN + FP}$$

CHAPTER- 2 EXPLORATORY DATA ANALYSIS

Exploratory data analysis (EDA) is a statistical approach that aims at discovering and summarizing a dataset. It is the first step in any data analysis process. Here, we make sense of the data we have and then figure out what questions we want to ask and how to frame them, as well as how best to manipulate our available data sources to get the answers we need.

We do this by taking a broad look at patterns, trends, outliers, unexpected results and so on in our existing data, using visual and quantitative methods to get a sense of the story this tells. We look for clues that suggest our logical next steps, questions or areas of research.

Text Preprocessing:

Text can come in a variety of forms from a list of individual words, to sentences to multiple paragraphs with special characters (like tweets for example). Like any data science problem, we understand the questions that are being asked will inform what steps may be employed to transform words into numerical features that work with machine learning algorithms.

Pre-processing the data is the process of cleaning and preparing the text for classification. Online texts contain usually lots of noise and uninformative parts such as HTML tags, scripts and advertisements. In addition, on words level, many words in the text do not have an impact on the general orientation of it. Keeping those words makes the dimensionality of the problem high and hence the classification more difficult since each word in the text is treated as one dimension. Here is the hypothesis of having the data properly pre-processed: to reduce the noise in the text should help improve the performance of the classifier and speed up the classification process, thus aiding in real time sentiment analysis.

The whole process involves several steps and we talk about them as we follow:

Converting all text to lowercase:

Here we convert all our text to lowercase. If there are 2 words “Restaurant” and “restaurant”, it will be converted to a single entry “restaurant”.

Removing Punctuations:

In this step we perform functions to remove punctuations and symbols such as !, %, $, etc.

Stopwords:

After processing with removal of punctuations, the next step in data preprocessing is removal of stopwords. A stop word is a commonly used word (such as "the") that are not relevant for doing analysis. So such words which are filtered out before or after processing of natural language data (text). In python we use ‘nltk’ package for removing stopwords. ‘nltk’ package already contains a list of stopwords, in addition to that we add our own custom list of stopwords they are words specific to the dataset that may not add value to the text and process it to the text to remove them.

Lemmatizing our text:

Stemming and Lemmatization are Text Normalization (or sometimes called Word Normalization) techniques in the field of Natural Language Processing that are used to prepare text, words, and documents for further processing. Stemming and Lemmatization have been studied, and algorithms have been developed in Computer Science since the 1960's. Stemming and Lemmatization helps us to achieve the root forms (sometimes called synonyms in search context) of inflected (derived) words. Stemming is different to Lemmatization in the approach it uses to produce root forms of words and the word produced.

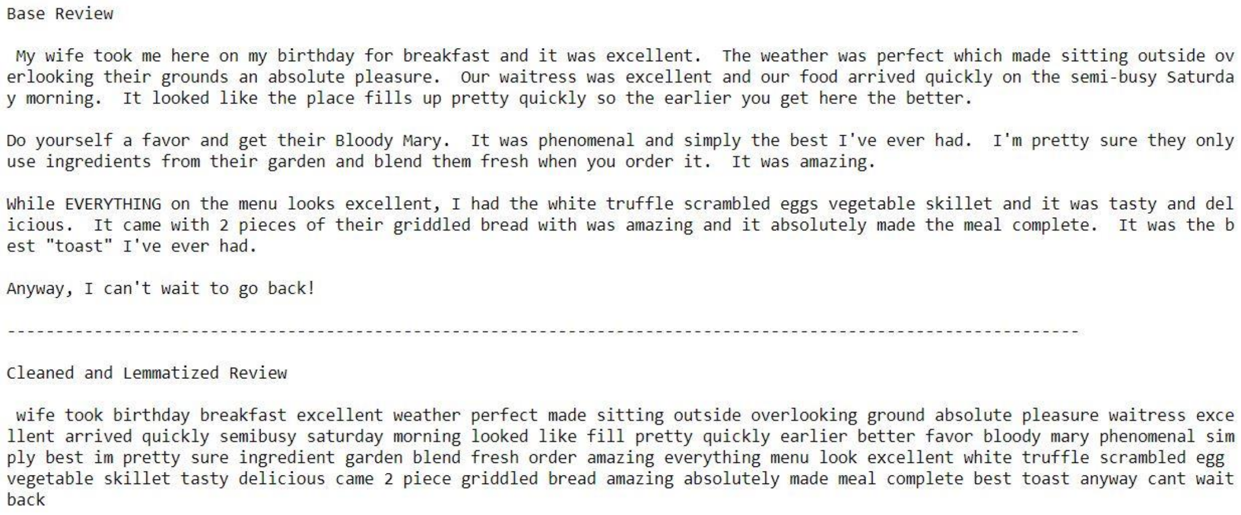
Stemming and Lemmatization are widely used in tagging systems, indexing, SEOs, Web search results, and information retrieval. For example, searching for fish on Google will also result in fishes, fishing as fish is the stem of both words.

Stemming algorithms work by cutting off the end or the beginning of the word, taking into account a list of common prefixes and suffixes that can be found in an inflected word. This indiscriminate cutting can be successful in some occasions, but not always, and that is why we affirm that this approach presents some limitations.

Lemmatization, on the other hand, takes into consideration the morphological analysis of the words.

Here in our analysis we do lemmatization of our text using ‘Word’ from textblob and ‘wordnet’ from nltk.

This is how one of our reviews looked before and after all the text pre-processing:



```
Base Review

 My wife took me here on my birthday for breakfast and it was excellent.  The weather was perfect which made sitting outside ov
erlooking their grounds an absolute pleasure.  Our waitress was excellent and our food arrived quickly on the semi-busy Saturda
y morning.  It looked like the place fills up pretty quickly so the earlier you get here the better.

Do yourself a favor and get their Bloody Mary.  It was phenomenal and simply the best I've ever had.  I'm pretty sure they only
use ingredients from their garden and blend them fresh when you order it.  It was amazing.

While EVERYTHING on the menu looks excellent, I had the white truffle scrambled eggs vegetable skillet and it was tasty and del
icious.  It came with 2 pieces of their griddled bread with was amazing and it absolutely made the meal complete.  It was the b
est "toast" I've ever had.

Anyway, I can't wait to go back!

-----------------------------------------------------------------------------------------------------------

Cleaned and Lemmatized Review

 wife took birthday breakfast excellent weather perfect made sitting outside overlooking ground absolute pleasure waitress exce
llent arrived quickly semibusy saturday morning looked like fill pretty quickly earlier better favor bloody mary phenomenal sim
ply best im pretty sure ingredient garden blend fresh order amazing everything menu look excellent white truffle scrambled egg
vegetable skillet tasty delicious came 2 piece griddled bread amazing absolutely made meal complete best toast anyway cant wait
back
```

Text Mining Approaches:

In text mining approaches, there are 2 basic categories:

- *Semantic Parsing* where the word sequence, word usage as noun or verb, hierarchical word structure etc matters
- *Bag of Words* where all the words are analyzed as a single token and order does not matter.

Our project will only be limited to the “bag of words” approach. So we move forward with Bag of words text mining approach.

A bag-of-words model, or BoW for short, is a way of extracting features from text for use in modeling, such as with machine learning algorithms. The approach is very simple and flexible, and can be used in a myriad of ways for extracting features from documents.

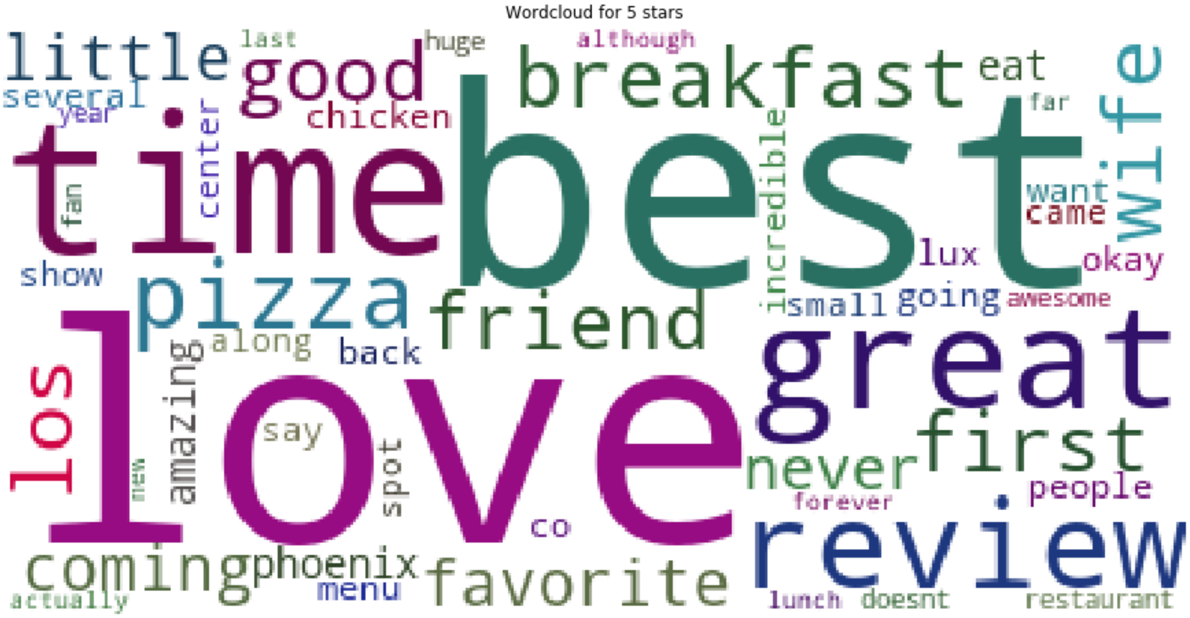
A bag-of-words is a representation of text that describes the occurrence of words within a document. It involves two things:

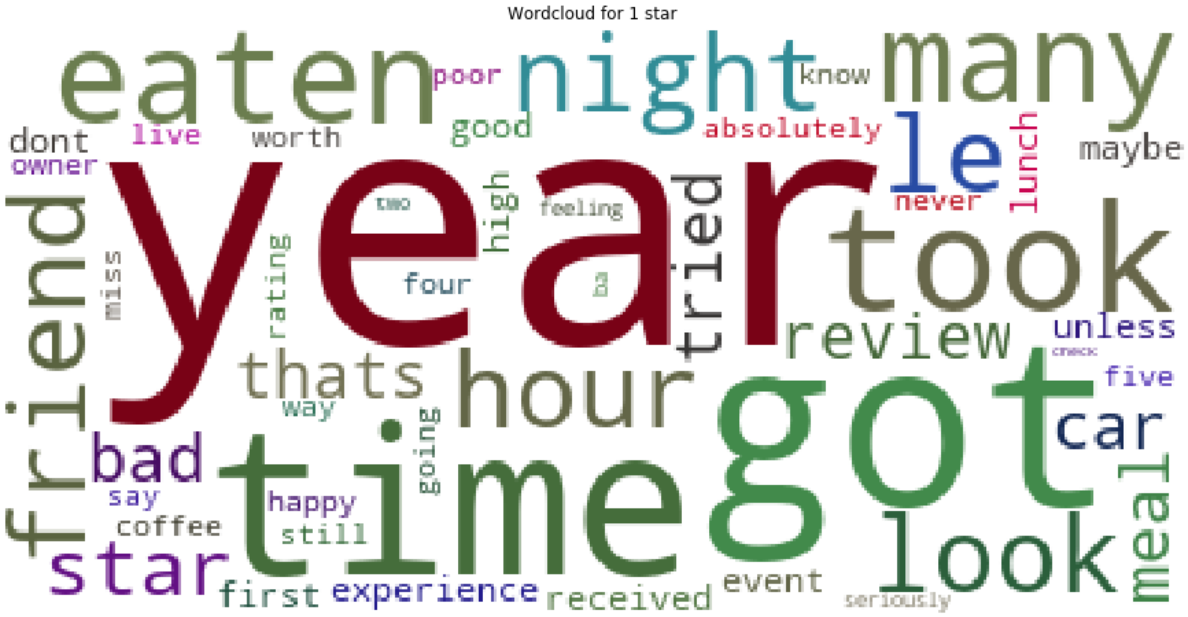
1. A vocabulary of known words.
2. A measure of the presence of known words.

It is called a “*bag*” of words, because any information about the order or structure of words in the document is discarded. The model is only concerned with whether known words occur in the document, not where in the document.

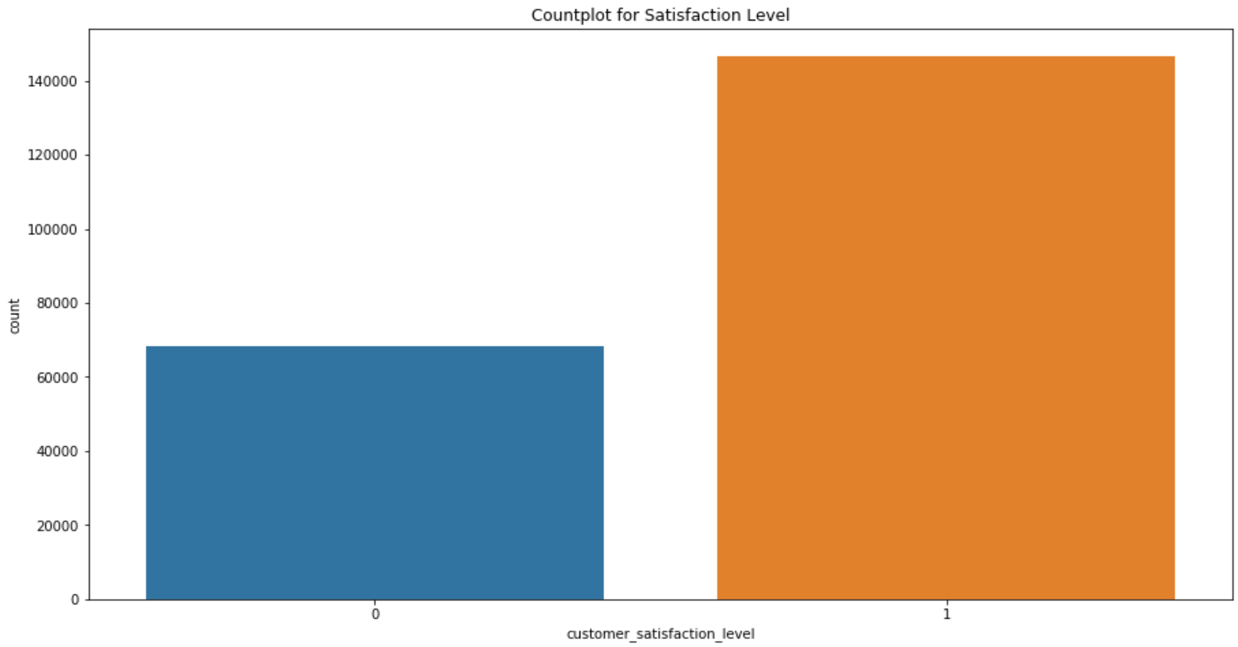
Visualizations and insights from various plots:

Exploratory Text Analysis



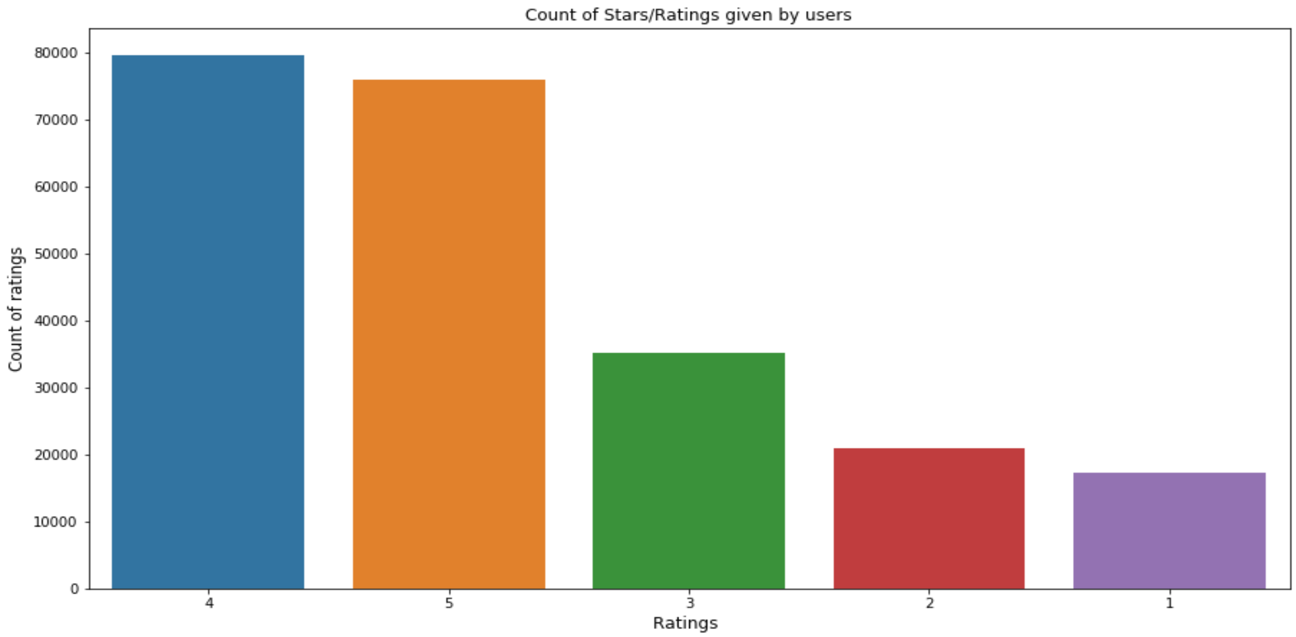


Visualizing the distribution of our target variable



- ✓ The plot is a countplot to see the distribution of our target variable which is a binary class variable with values labeled as 0 and 1, where 0 implies unhappy customers and 1 implies happy customers.
- ✓ From the graph it is pretty clear that our data is imbalanced with 68% towards the positive class (1) and 32% towards the negative class (0).

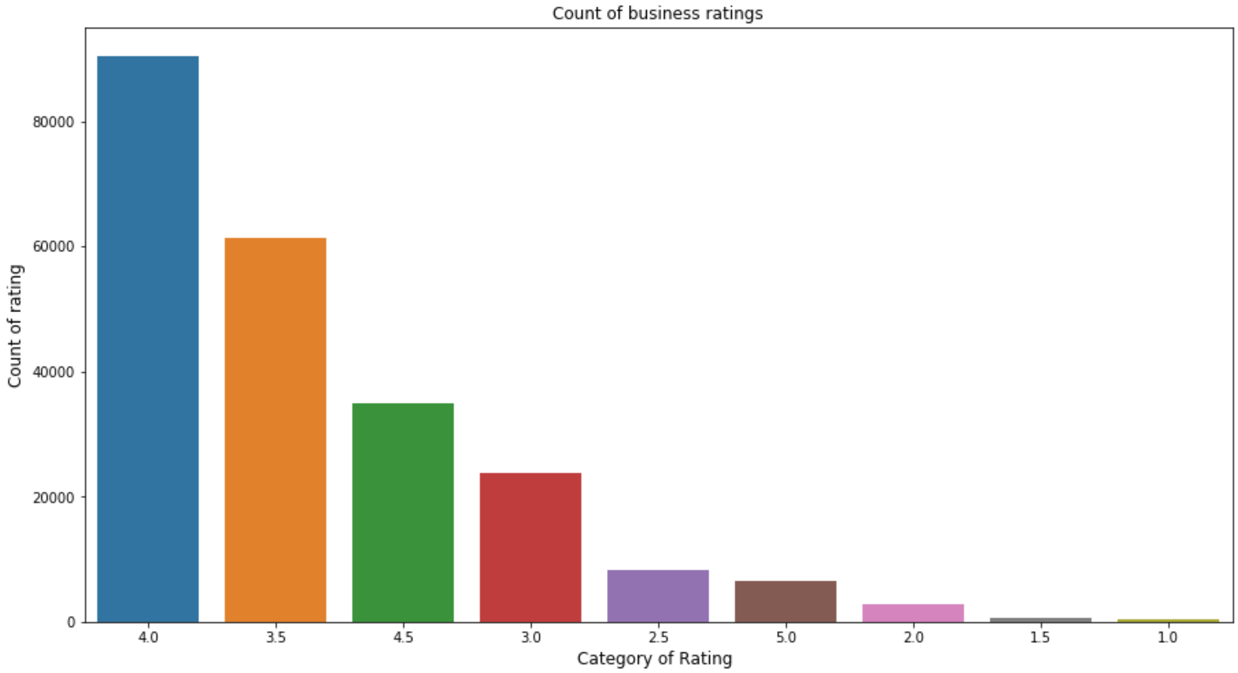
Visualizing the count of star distribution rated by users



- ✓ From graph we can see that 4 star ratings are the maximum in count (close to 80000) as rated by the users followed by 5 star ratings in the 2nd place and not much less in count as that of 4 star ratings.

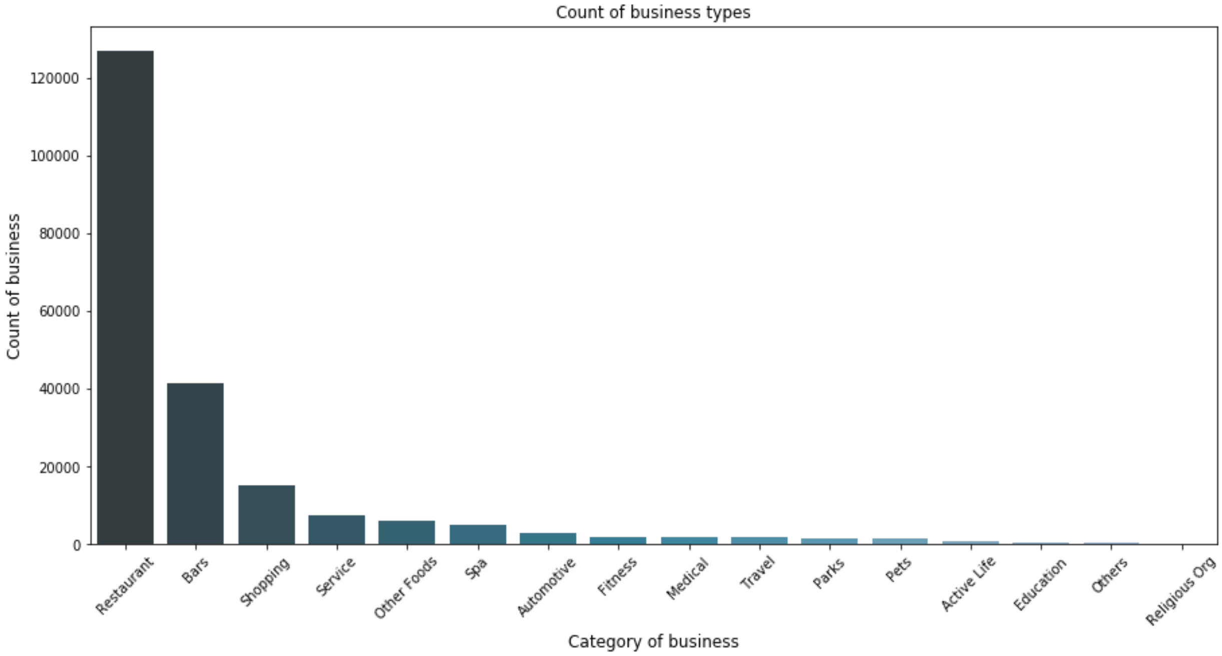
- ✓ The least number of counts is given to 1 star class followed by 2 star which is actually a good sign that customers are more satisfied as compared to dissatisfaction.

Visualizing the count of business ratings



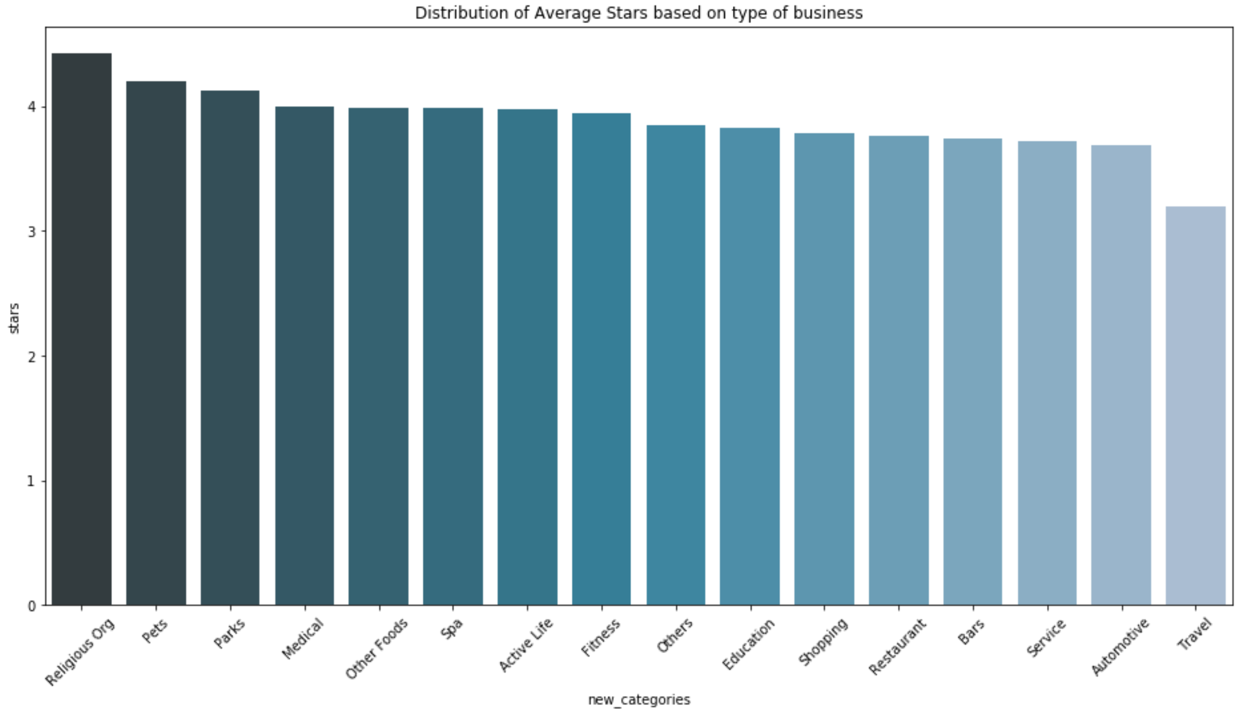
- ✓ This is a plot to see the variation of count of business stars. Though it is not specified to us in the dataset, but the information is being given to us as a column named ‘business_stars’.
- ✓ Here from our plot we can conclude that most of the businesses are rated 4 and then a drop in count near about 20000 which is 3.5 rating.

Visualizing the count of business categories



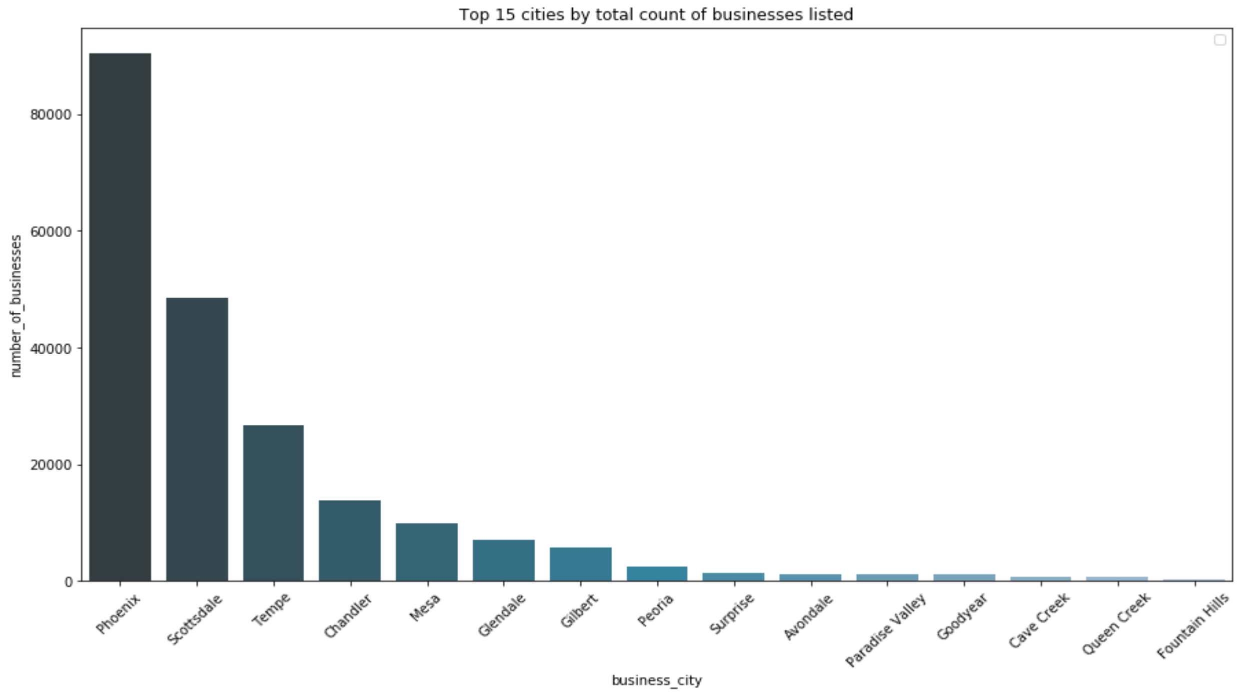
- ✓ We have classified our “business_categories” attribute to 16 major categories based on the key words in them. The above plot is the count of them.
- ✓ Restaurants being the majority in number with more than 130000 and religious organizations being the least.

Distribution of business categories based on average stars



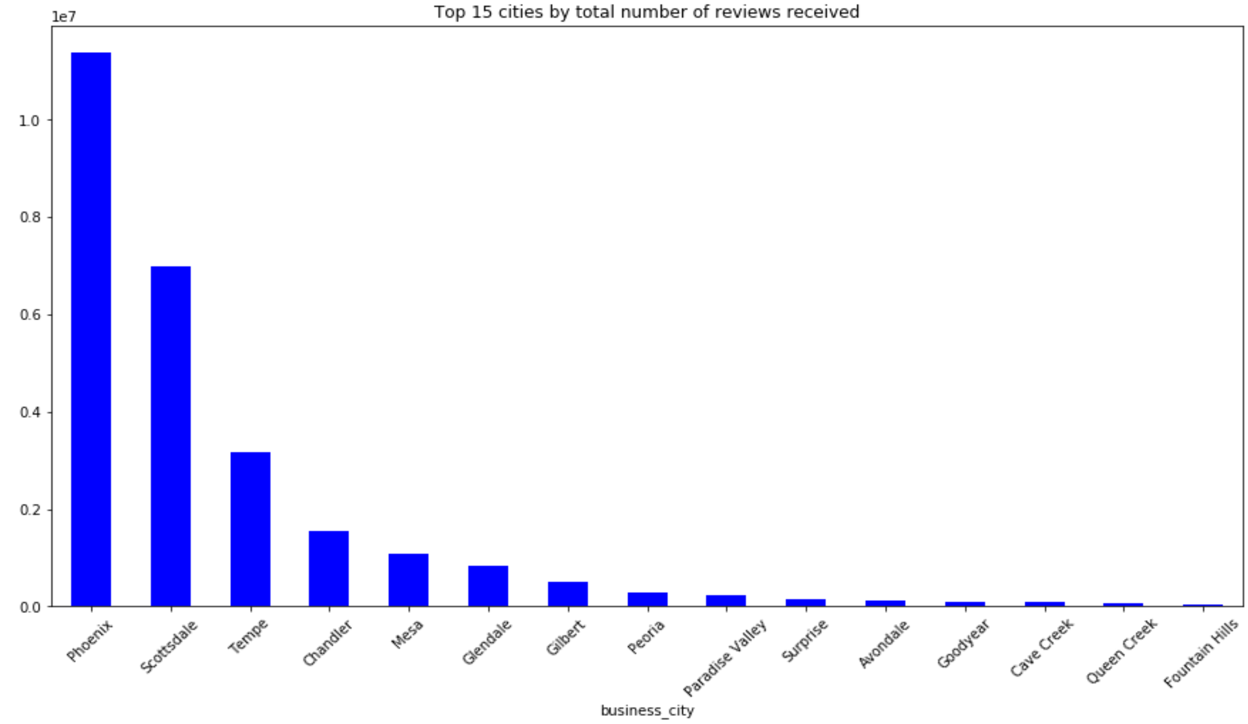
- ✓ This plot explains the average stars received based on the newly created business categories.
- ✓ Maximum being achieved by religious organizations and least by travel
- ✓ But a noteworthy point to look at is that there is no category with average rating below 3which is really good.

Top 15 cities by businesses listed



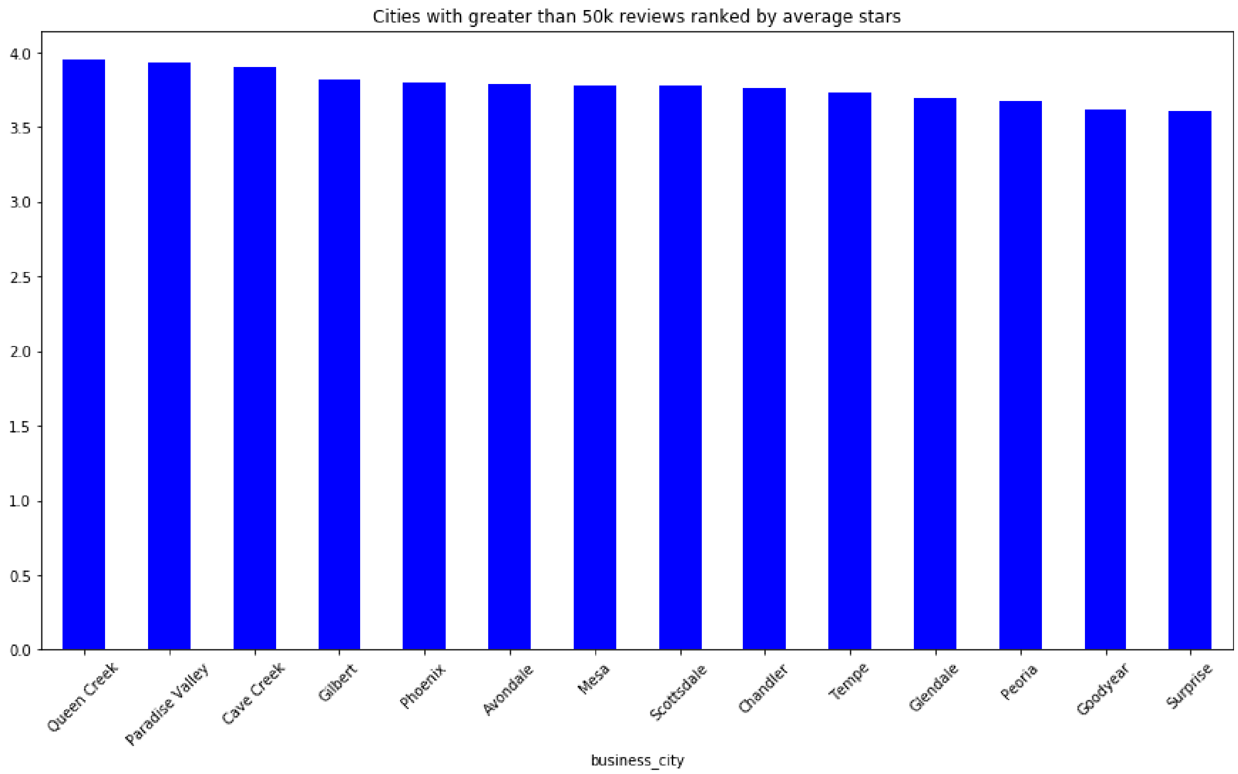
- ✓ The distribution above implies the total count of top 15 cities by business.
- ✓ Phoenix has emerged to be the top city with the maximum number of business with more than 80000 businesses.
- ✓ Scottsdale bags the 2nd spot with near about 50000 businesses.

Top 15 cities based on reviews



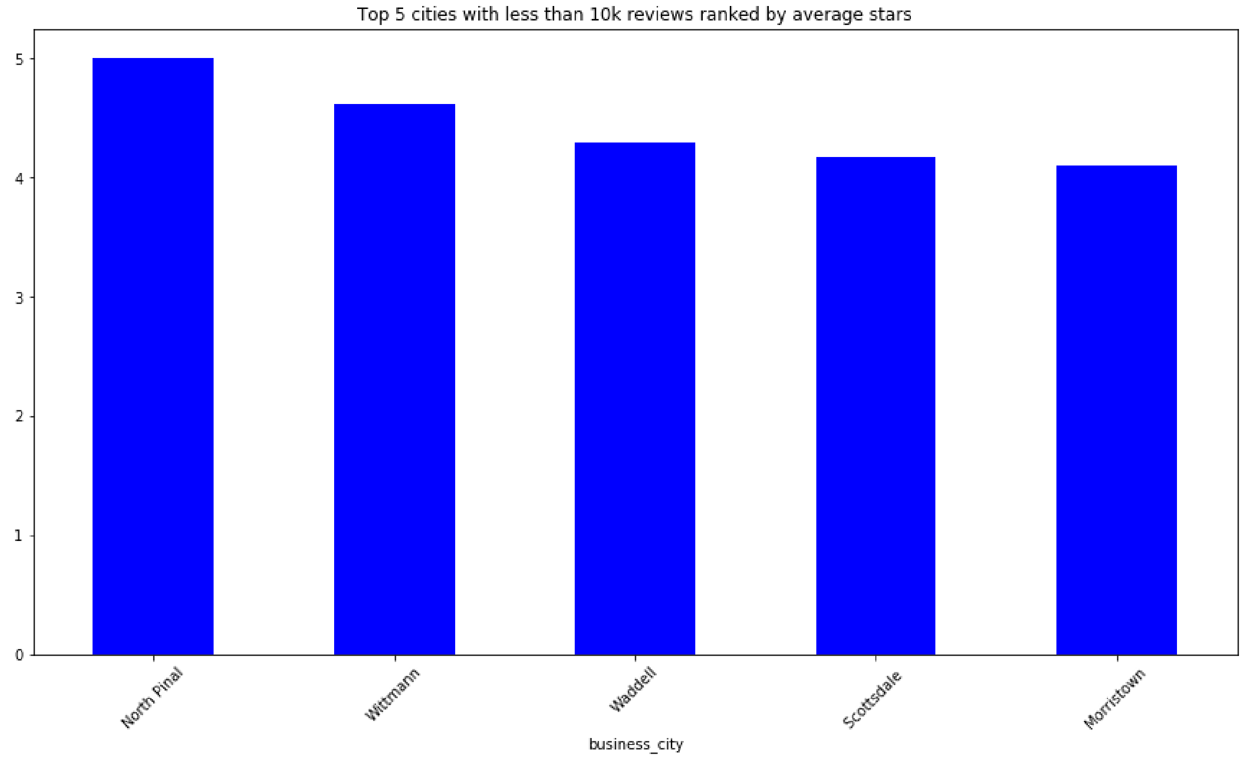
- ✓ Based on the total number of reviews a city has received, we have plotted the above distribution containing the top 15 cities**.**
- ✓ Here also Phoenix occupies the top spot and followed by Scottsdale.

Cities with greater than 50k reviews ranked by average stars



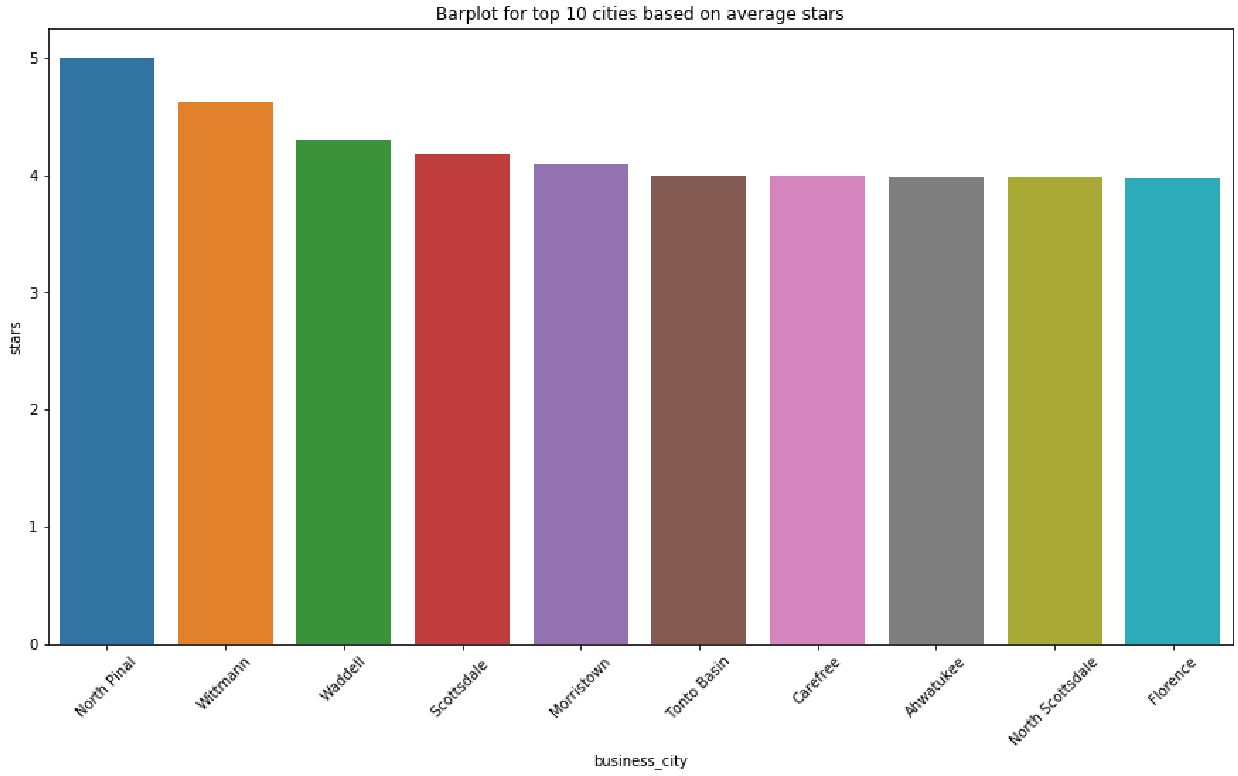
- ✓ From this graph we can see that cities with greater than 50000 reviews have a average rating of more than 3.5
- ✓ Since most reviewed cities got an average above 3.5 so almost all business in these cities are satisfying all customers
- ✓ But this cannot be treated as a positive conclusion as the range lies between 3.5 to 4.

Top 5 cities with less than 10k reviews ranked by average stars



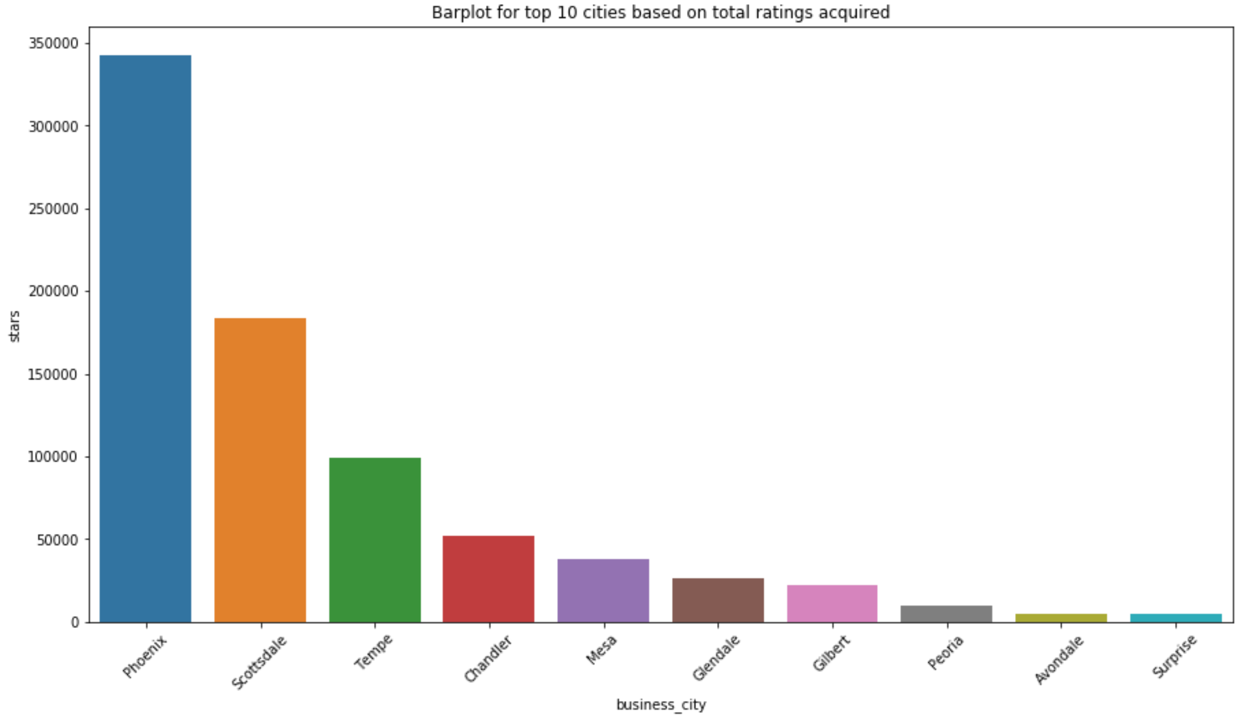
- ✓ From the above plot we can see that cities with less than 10000 reviews have a average rating of more than 4
- ✓ The range of average stars lies between 4 and 5

Barplot for top 10 cities based on average stars



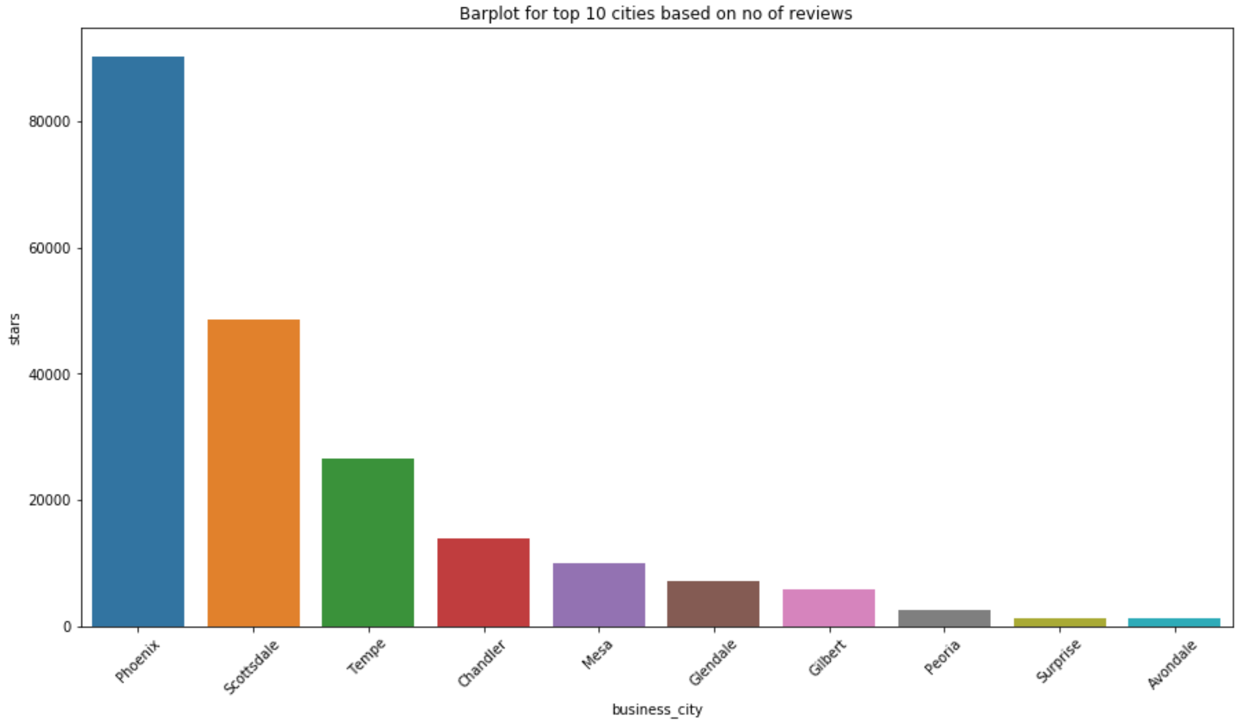
- ✓ The distribution gives us the distribution of top 10 cities based on average stars.
- ✓ North Pinal tops in cities receiving the highest average number of stars
- ✓ The cities occupying 6th, 7th ,8th ,9th and 10th position have almost similar number of average stars

Checking the trend of top 10 cities based on the total number of stars



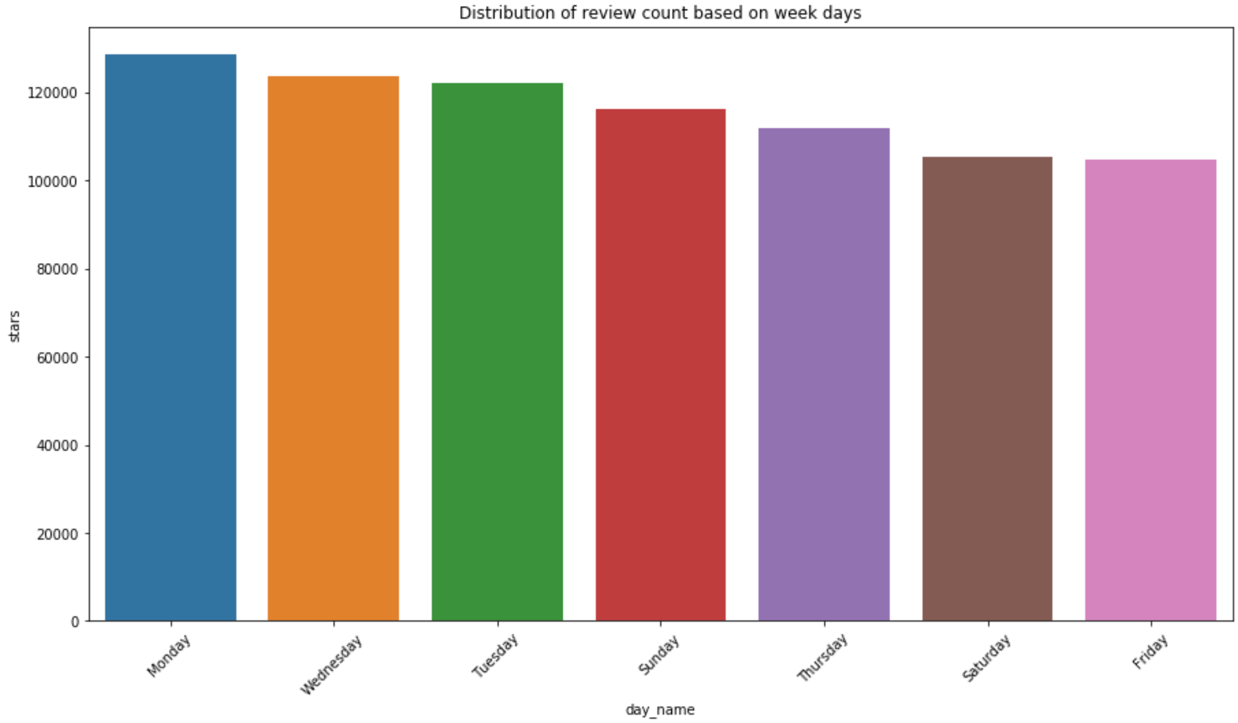
- ✓ The above plot gives us the top 10 cities based on the sum of total number of reviewer stars.
- ✓ Phoenix attained the top spot with total stars close to 35000.

Checking the trend of top 10 cities based on no of reviews



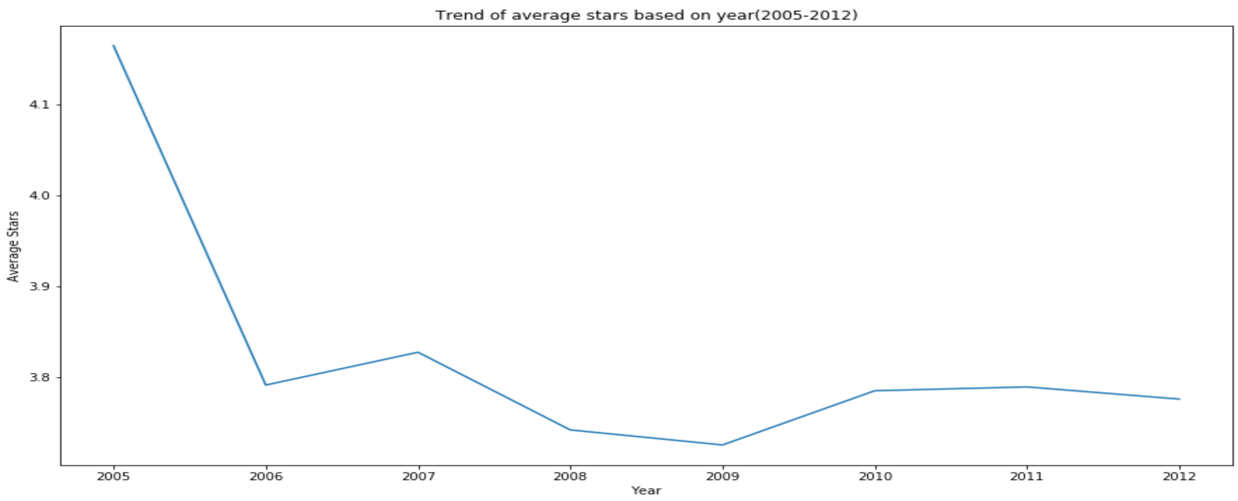
- ✓ The plot above gives us the top 10 cities based on the number of reviews each city has received.
- ✓ Even here Phoenix emerged as the top city with max number of reviews received.

Checking the trend of review count based on days of a week



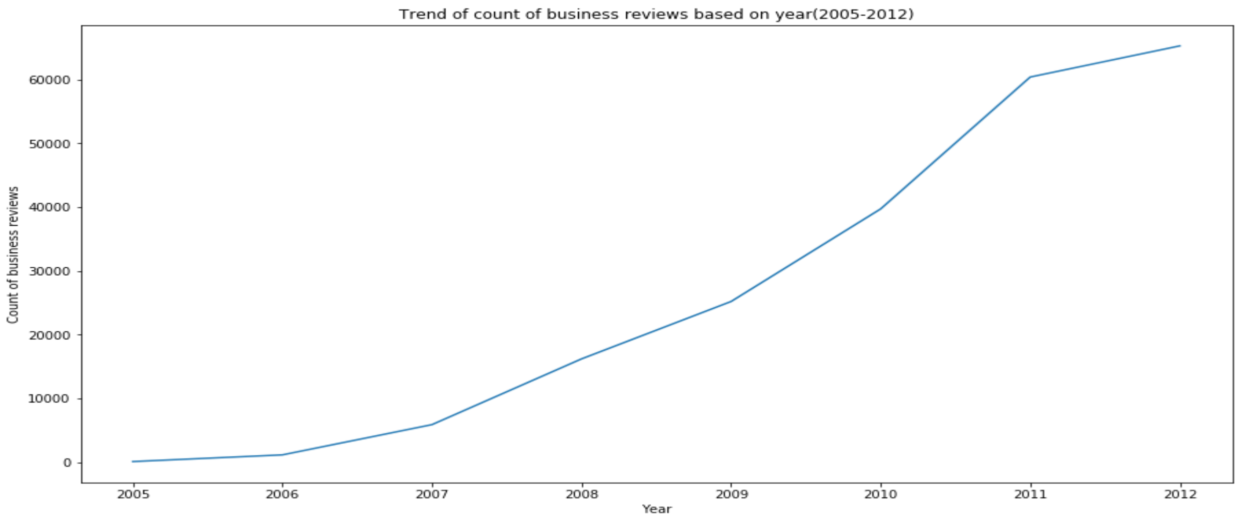
- ✓ The above distribution is to understand the relationship between days of a week and the count of reviews.
- ✓ From plot we can find Monday and Tuesday people are giving more reviews which doesn’t directly mean that on Monday and Tuesday more people visit these businesses.
- ✓ Friday and Saturday number of reviews are less

Checking the trend of review stars(average) based on year



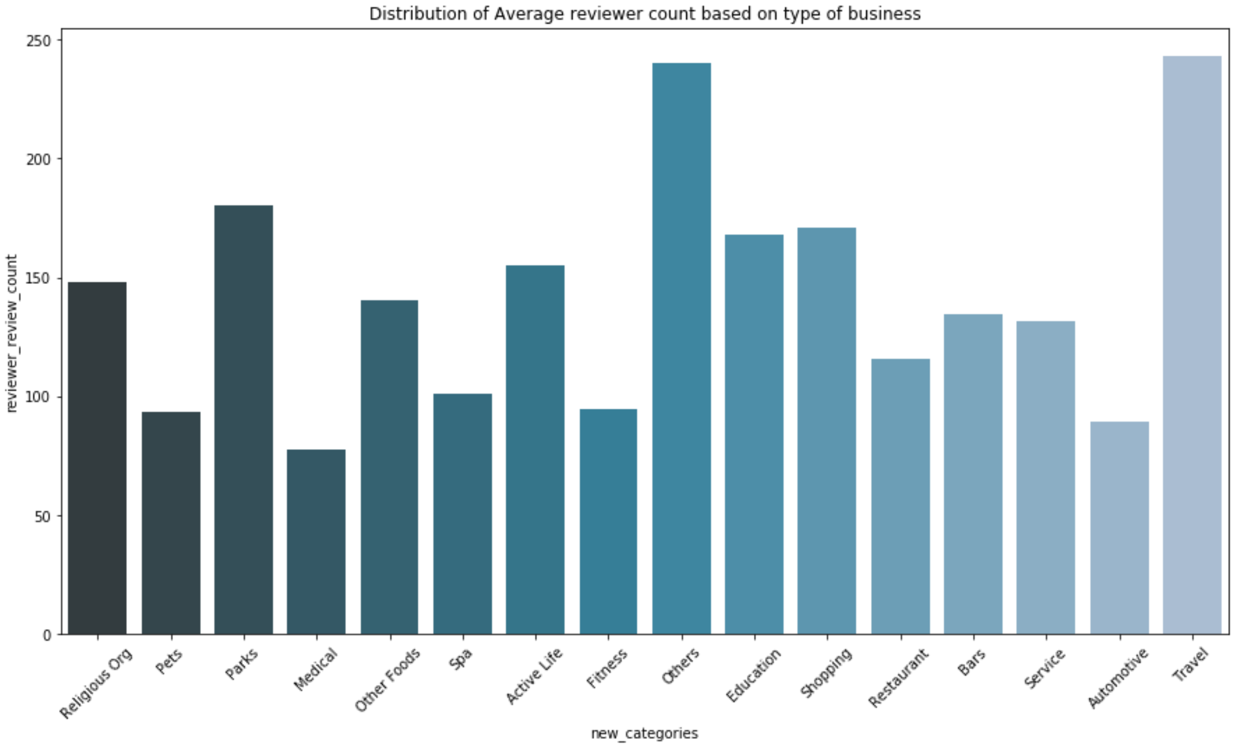
- ✓ Average rating in 2005 is above 4.1
- ✓ From graph we can see sudden drop in rating from 2005 to 2006
- ✓ After 2006 variation in rating is not very high

Looking the trend of review count based on year



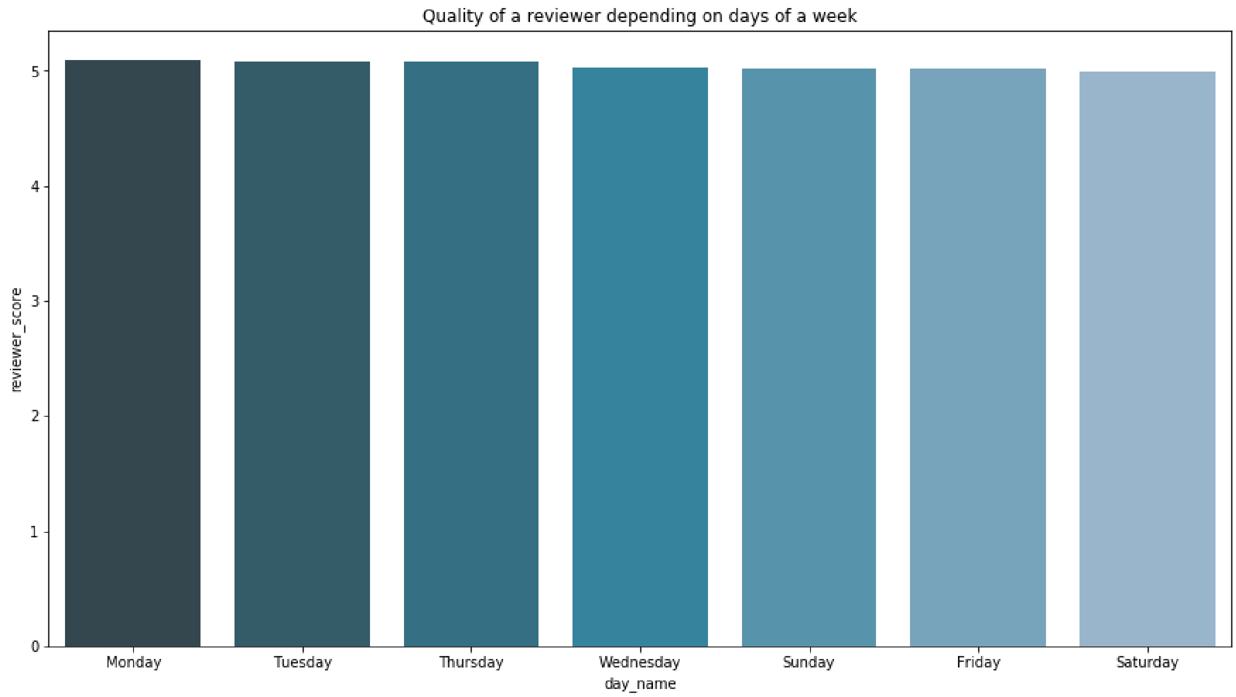
- ✓ From graph we can see that number of review count is increasing from 2005 to 2011 at a really good rate.
- ✓ Whereas from 2011 to 2012, though the count is increasing, the rate has declined a little and is gaining a nearly linear growth.

Distribution of Average reviewer count based on type of business



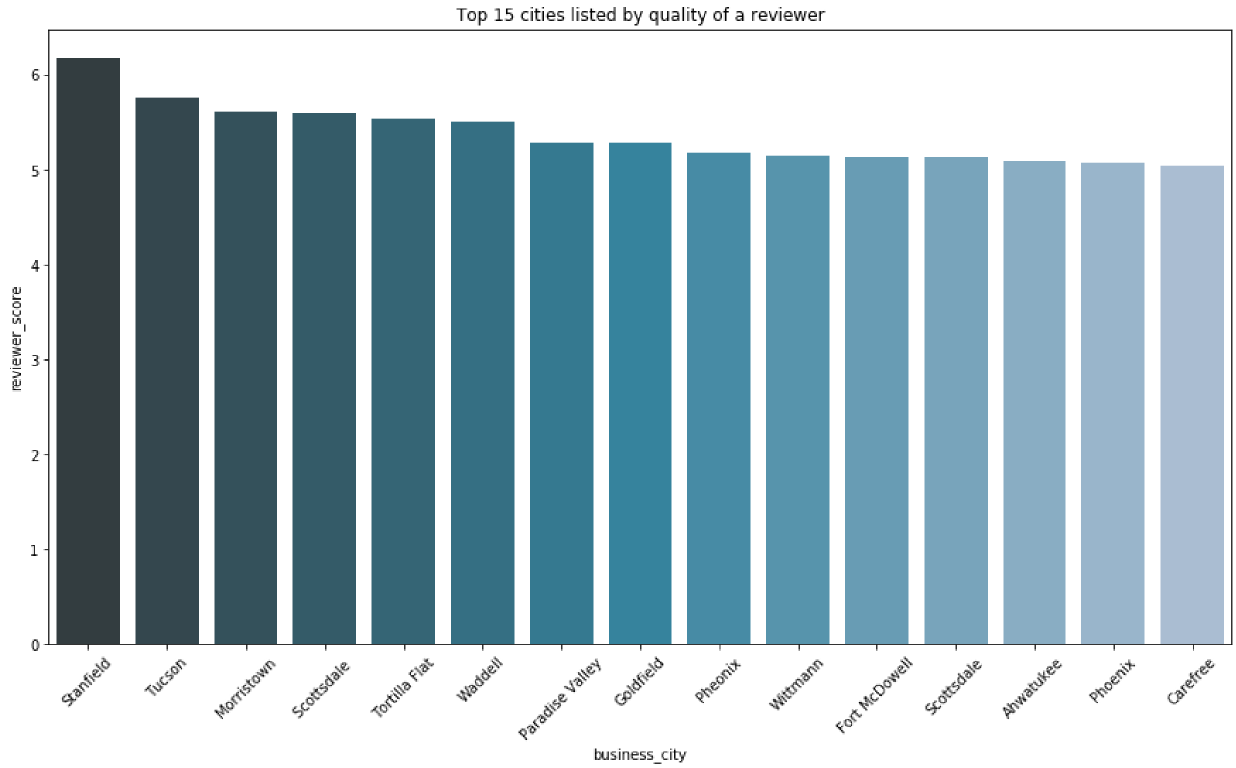
- ✓ The above plot is to understand the average distribution of stars based on business type.
- ✓ Travel category of business has topped in gaining average stars which means that people love to travel and also that service of this type of business is the best among the other categories.
- ✓ The last spot is bagged by medical, meaning patients are not satisfied with the type of service they are getting from medical facilities.

Distribution of days of a week listed by quality of reviewers



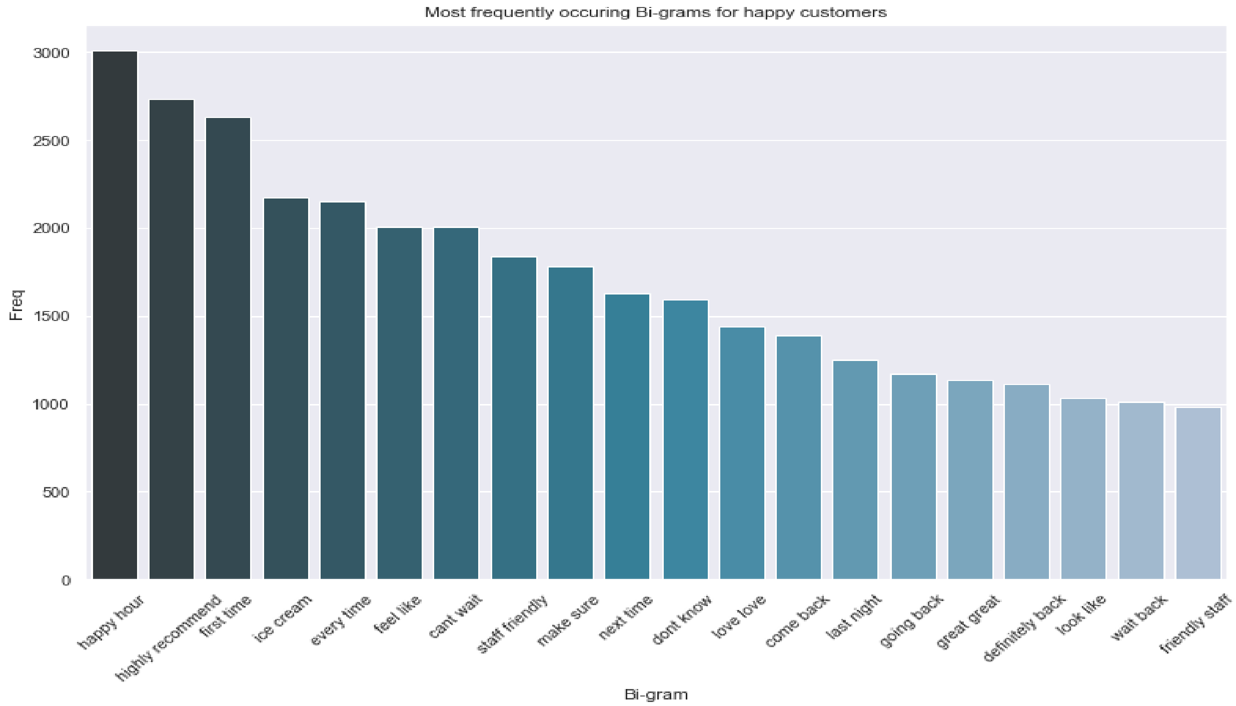
- ✓ The above plot lists the days of a week in descending order of quality of reviewer.
- ✓ The interpretation of the graph is done as Monday occupying the top spot with average reviewer score, which means that on Monday most of the customers are more trustworthy when compared to the other days of a week.
- ✓ This is just to explain how to interpret the above graph, in a different and more meaningful note it is seen that the score is pretty much the same across all the 7 days a week.

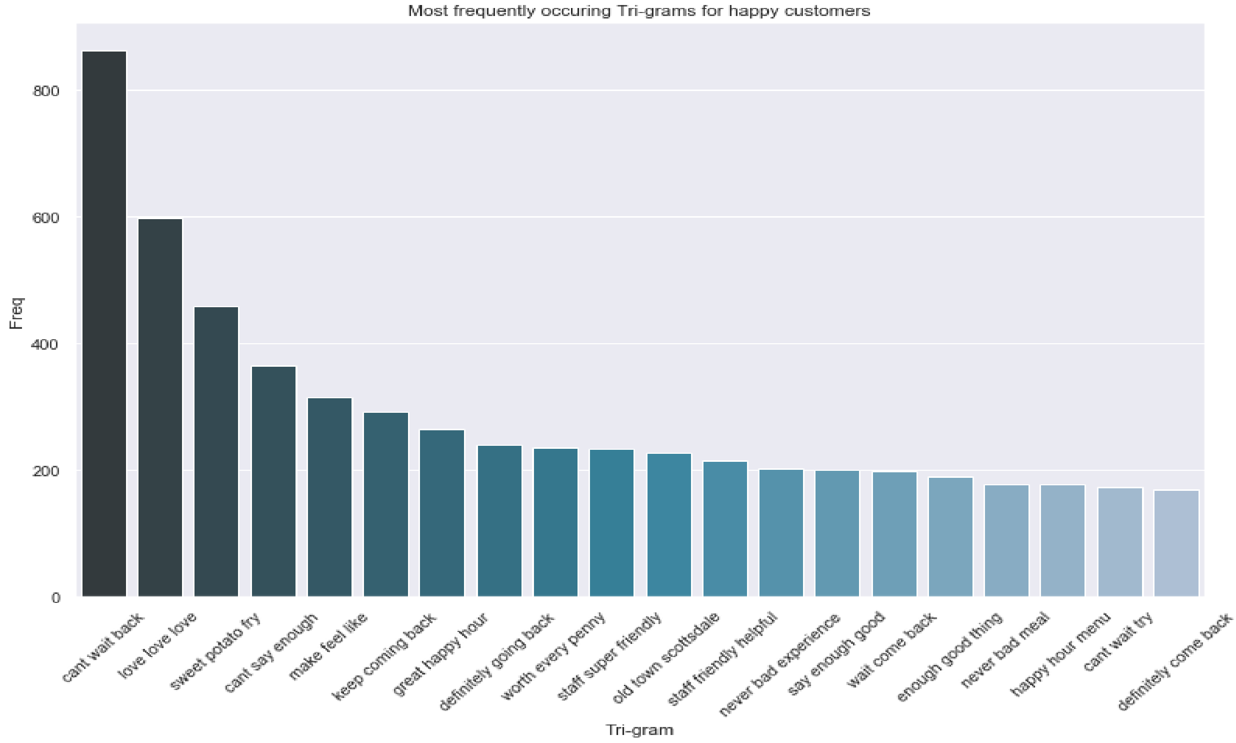
Distribution of top 15 cities listed by quality of reviewer



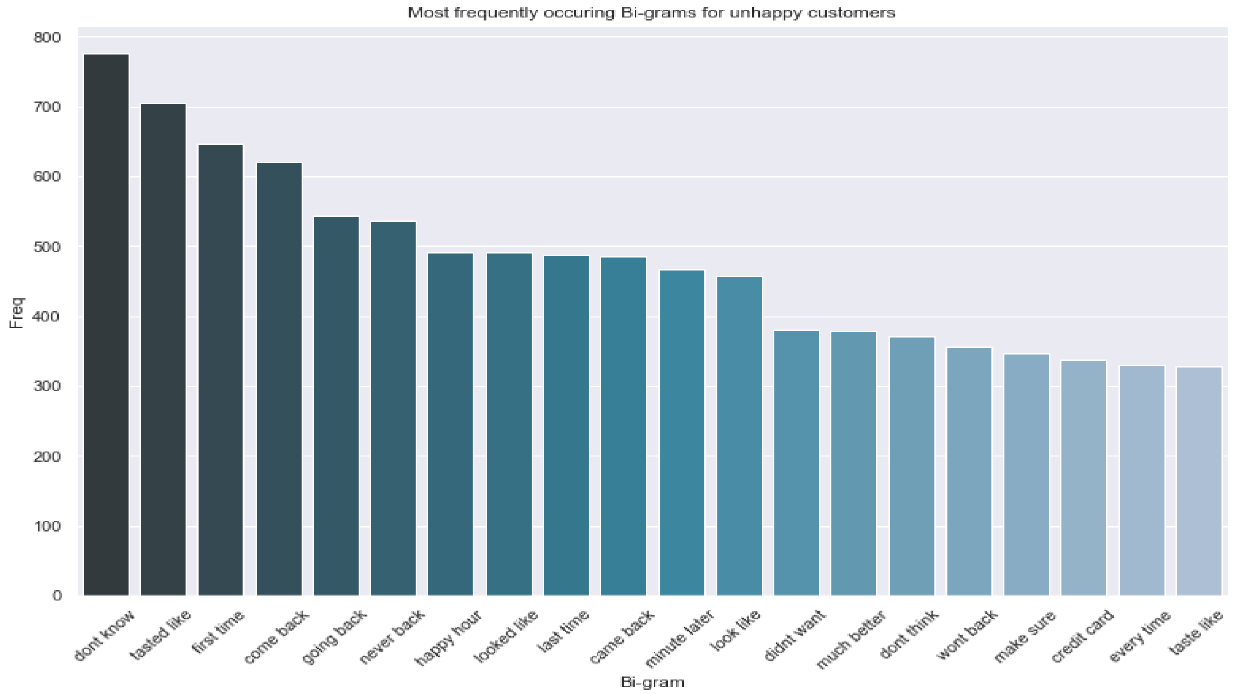
- ✓ This graph gives us a very good observation to understand our customer base.
- ✓ The distribution lists the top 15 cities based on average reviewer score.
- ✓ So if someone wants to understand and open a new business where one can consider the feedbacks and reviews on a true note, the above distribution gives the top 15 cities for the same.

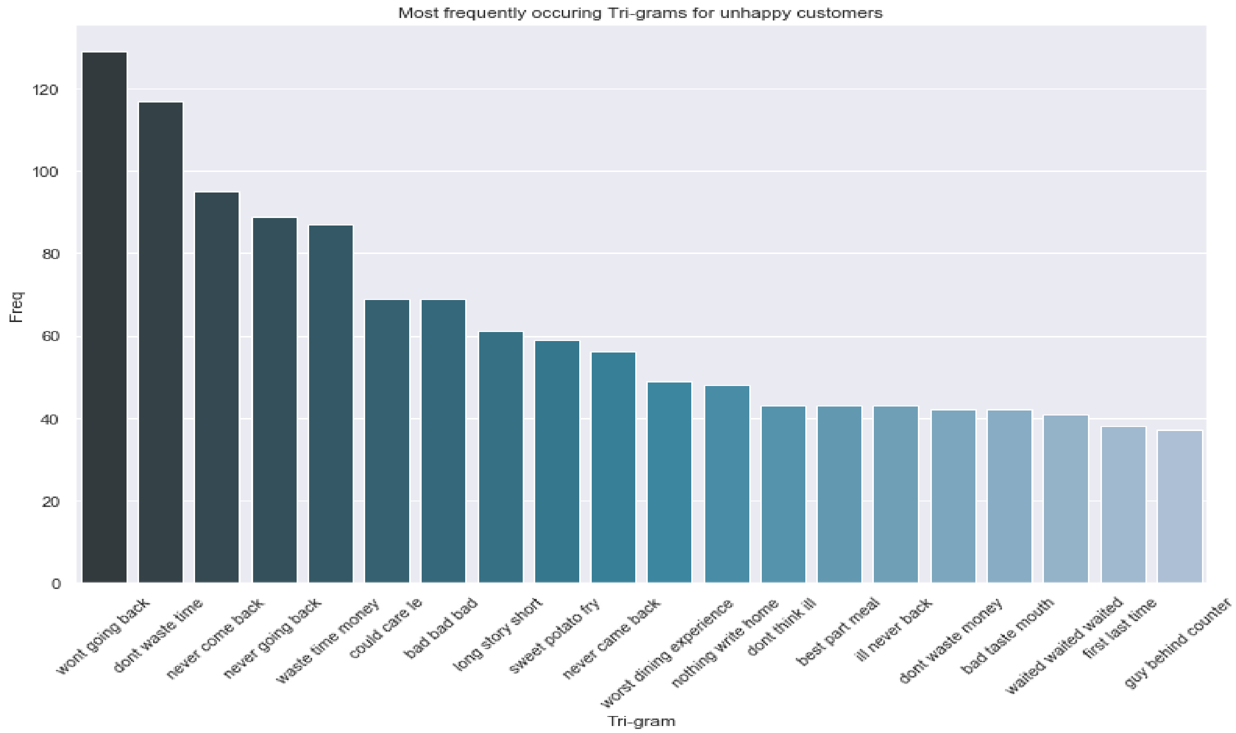
Most frequently occurring bi-grams and tri-grams for happy customers



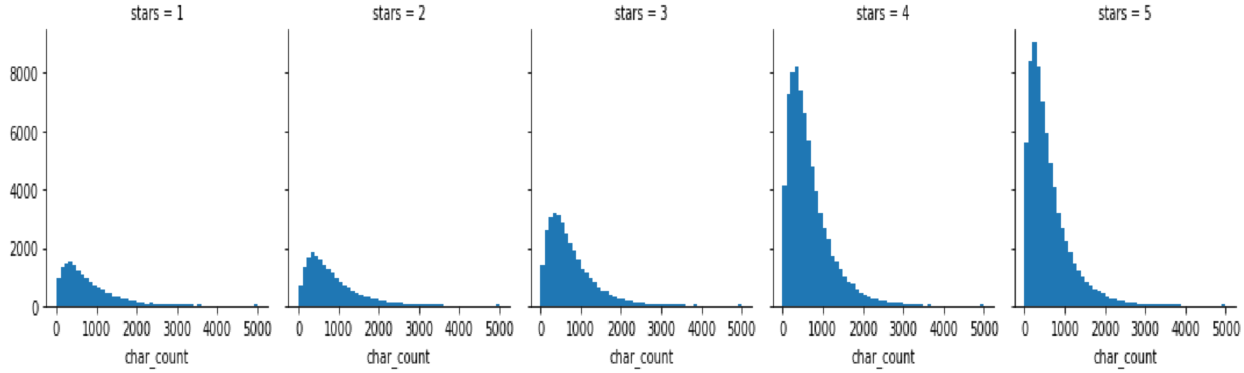


Most frequently occurring bi-grams and tri-grams for unhappy customers



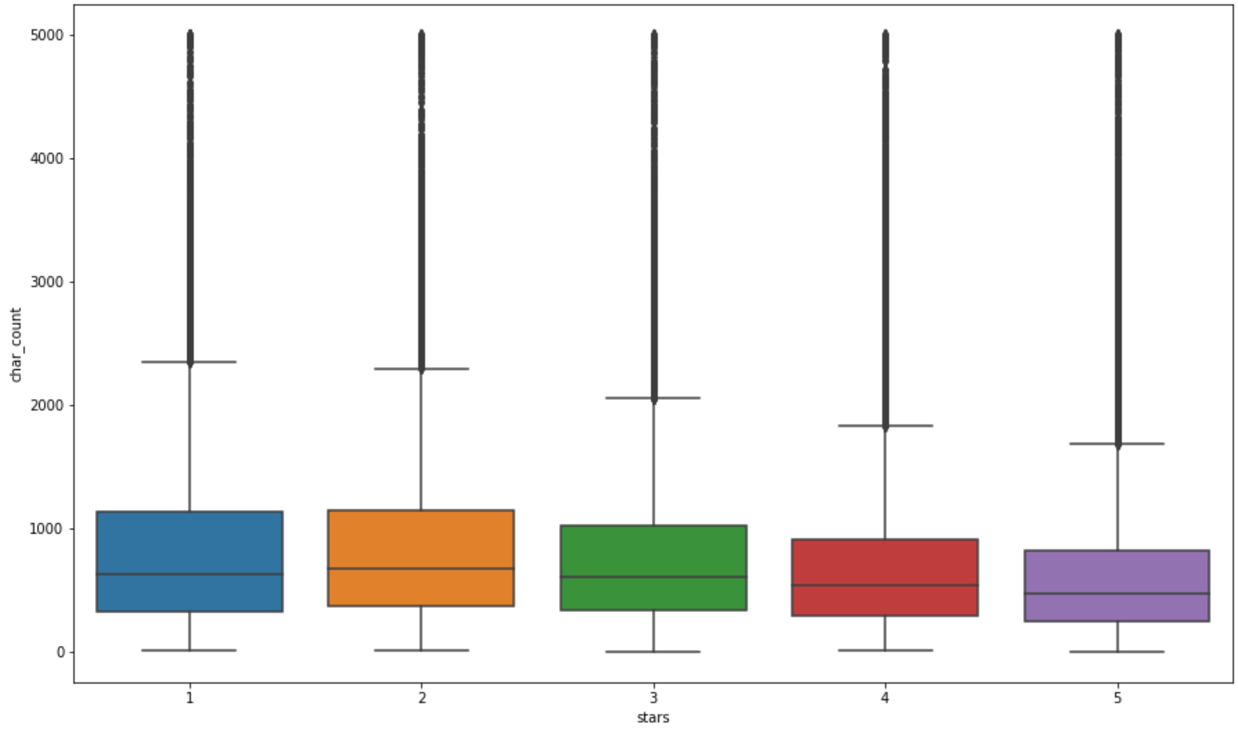


Distribution of character length of each review using Seaborn’s “FacetGrid”



- ✓ Seaborn’s FacetGrid allows us to create a grid of histograms placed side by side. We can use FacetGrid to see if there’s any relationship between our newly created 'char_count' feature and the stars rating.
- ✓ We have analysed the character length of each review to see if the newly created column of character length gives us any meaningful insights which might help us in our analysis.
- ✓ The distribution of text length is similar across all five ratings. However, the number of text reviews seems to be skewed a lot higher towards the 4-star and 5-star ratings.
- ✓ The skew in the distributions won’t help us, rather it might complicate our analysis.
- ✓ So we decided not to use this feature for our analysis.

Distribution of character length for each star rating



- ✓ The above plot is a box-plot for character length of each star rating.
- ✓ From the plot, looks like the 1-star and 2-star ratings have much longer text, but there are many outliers (which can be seen as points above the boxes).
- ✓ Because of this, maybe text length won't be such a useful feature to consider after all.

CHAPTER- 3 LOGISTIC REGRESSION

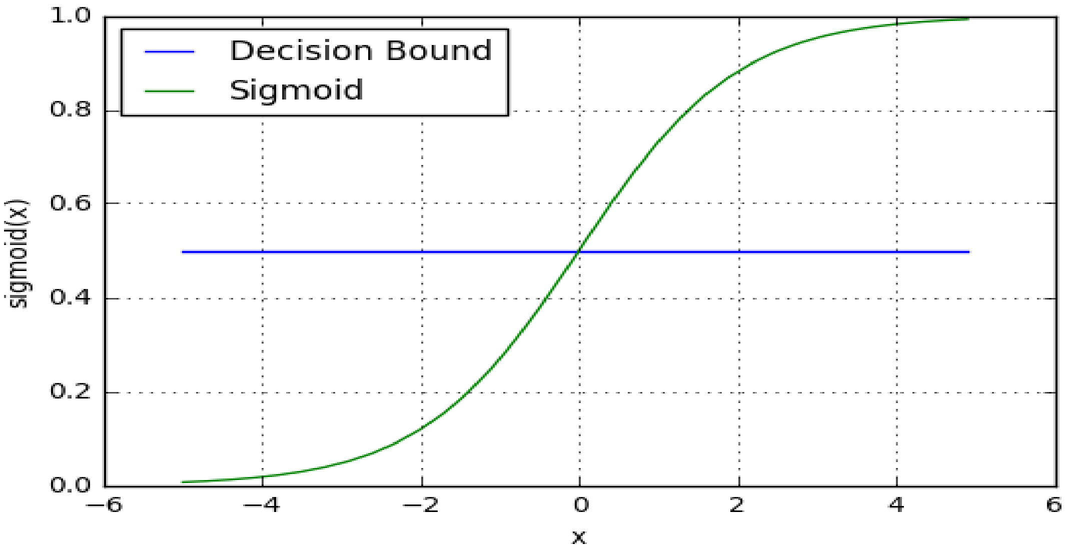
Logistic regression classifier is more like a **linear classifier** which uses the calculated logits (score) to predict the target class. It is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

Below is the most accurate and well-defined definition of logistic regression from Wikipedia:

"Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a ***logistic function****"*

The dependent and the independent variables are the same which we use in building a simple linear regression model. The dependent variable is the target class variable ("stars" in our problem) we are going to predict. However, the independent variables are the features or attributes we are going to use to predict the target class ("text" in our case).

Unlike in linear regression where given the value of an independent variable, we can predict our target value at instance of that independent variable. Whereas in the case of Logistic Regression, we have to map the predicted values to probabilities. In order to map these predicted values to probabilities, we use the sigmoid function. The function maps any real value into another value between 0 and 1. In machine learning, we use sigmoid to map predictions to probabilities.



The above graph is nothing but the Sigmoid Graph which is plotted by using the Sigmoid Function.

Our current prediction function returns a probability score between 0 and 1. In order to map this to a discrete class (true/false, 0/1), we select a threshold value or tipping point above which we will classify values into class 1 and below which we classify values into class 0.

$$P \geq 0.5\ ,\ \text{class} = 1$$

$$P < 0.5\ ,\ \text{class} = 0$$

For example, if our threshold was .5 and our prediction function returned .7, we would classify this observation as class 1. If our prediction was .2 we would classify the observation as class 0. For logistic regression with multiple classes we could select the class with the highest predicted probability.

Relevance of Logistic regression for our dataset

In our yelp dataset for modeling first we are creating a model for predicting if a customer is satisfied or dissatisfied based on the review one has given. What we did in our analysis is that, there are 5 classes of ratings a user can give to a business. The classes ranges from 1 being the least rating one can give up to 5 stars which is the best rating. So to make our analysis a binary classification problem, we clubbed the reviews which are rated 1, 2 and 3 as 'dissatisfied' and labeled it as 0 and we clubbed the remaining reviews rated 4 and 5 as 'satisfied' and labeled it as 1.

Logistic regression is used for predicting binary class problems .
We use logistic regression for predicting the class based on customer satisfaction as 0 or 1.

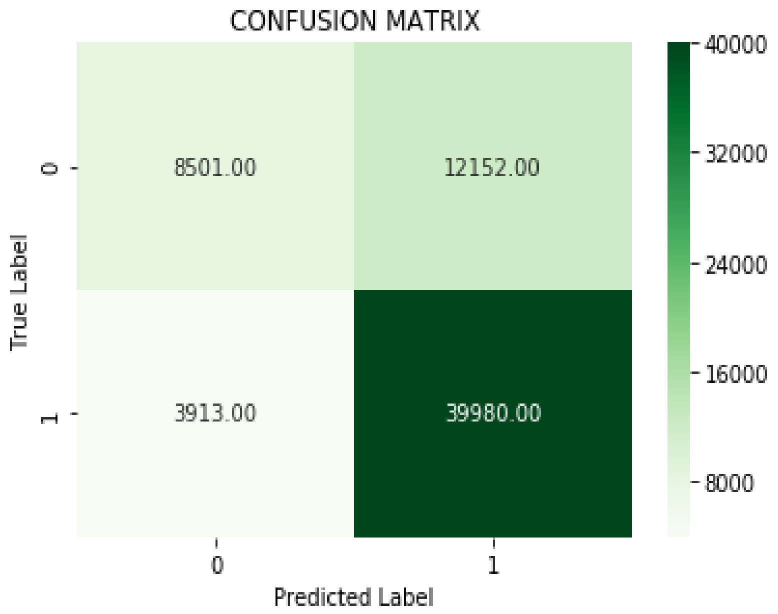
Logistic regression Summary

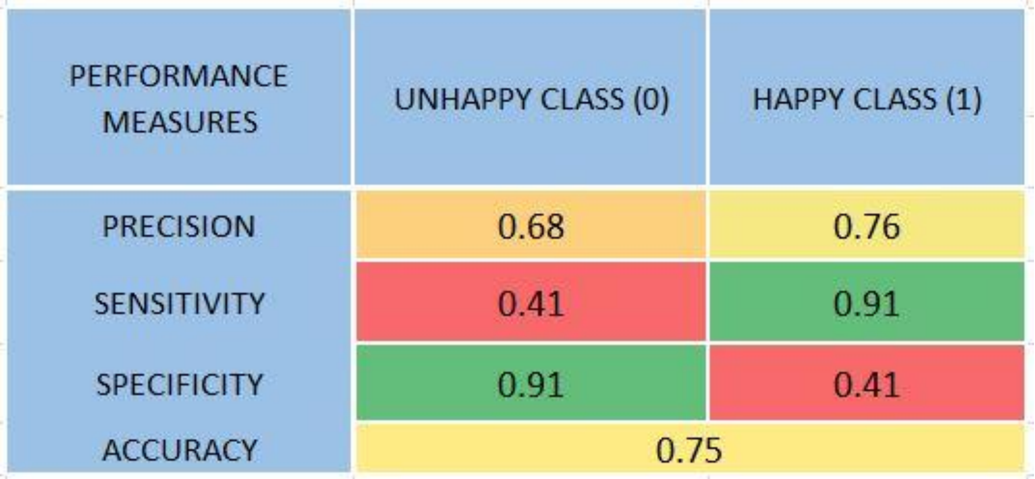
Equation:

logit(p) $= \log\left(\frac{p(y=1)}{1-(p=1)}\right) = \beta_0 + \beta_1 \cdot x_{i2} + \beta_2 \cdot x_{i2} + \_ + \beta_p \cdot x_{im}$

Where y is the dependent variable or target and $\beta_0 + \beta_1 \cdot x_{i2} + \beta_2 \cdot x_{i2} + \_ + \beta_p \cdot x_{im}$ are the independent or predictor variable

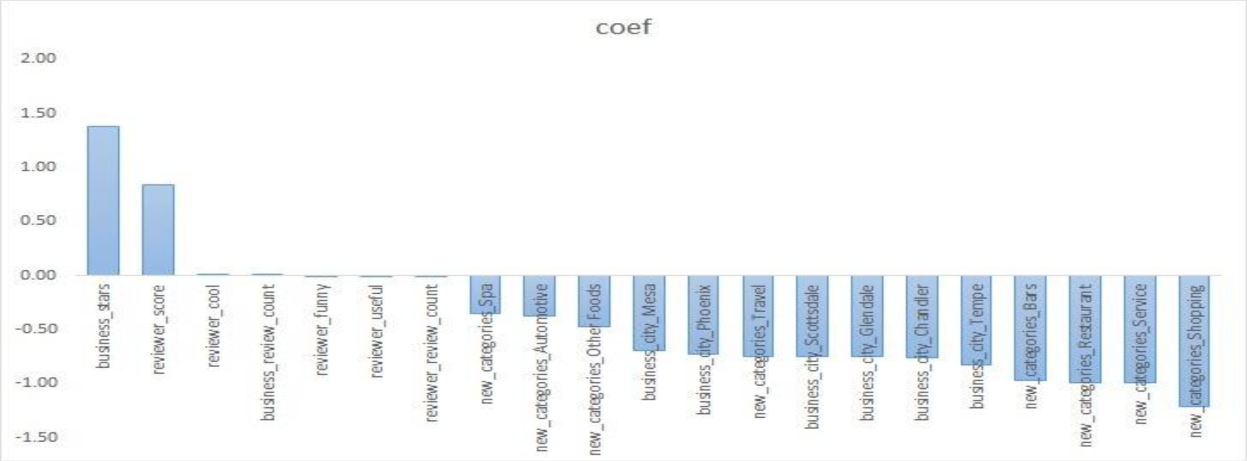
Confusion Matrix and Performance Measures for Structured Data Model

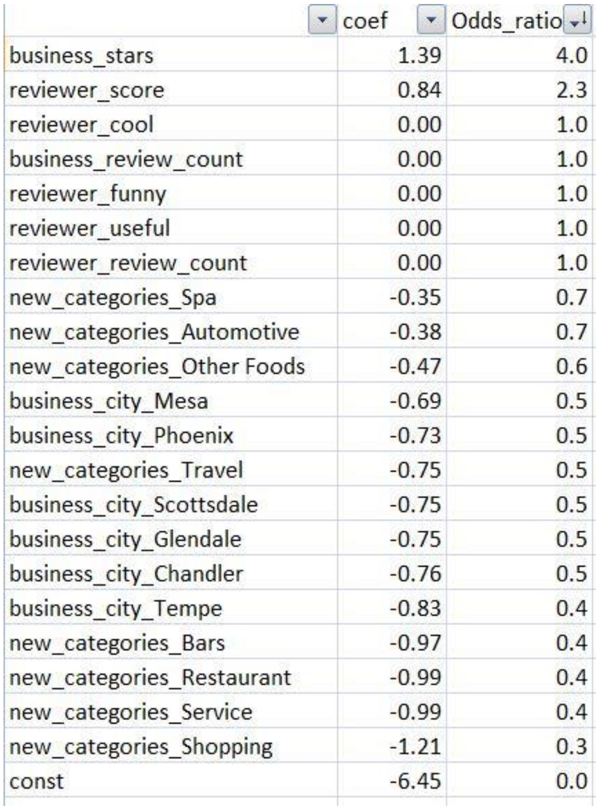




PERFORMANCE MEASURES	UNHAPPY CLASS (0)	HAPPY CLASS (1)
PRECISION	0.68	0.76
SENSITIVITY	0.41	0.91
SPECIFICITY	0.91	0.41
ACCURACY	0.75	

Co-efficient and odds ratio for our Logistic Model for Structured Data





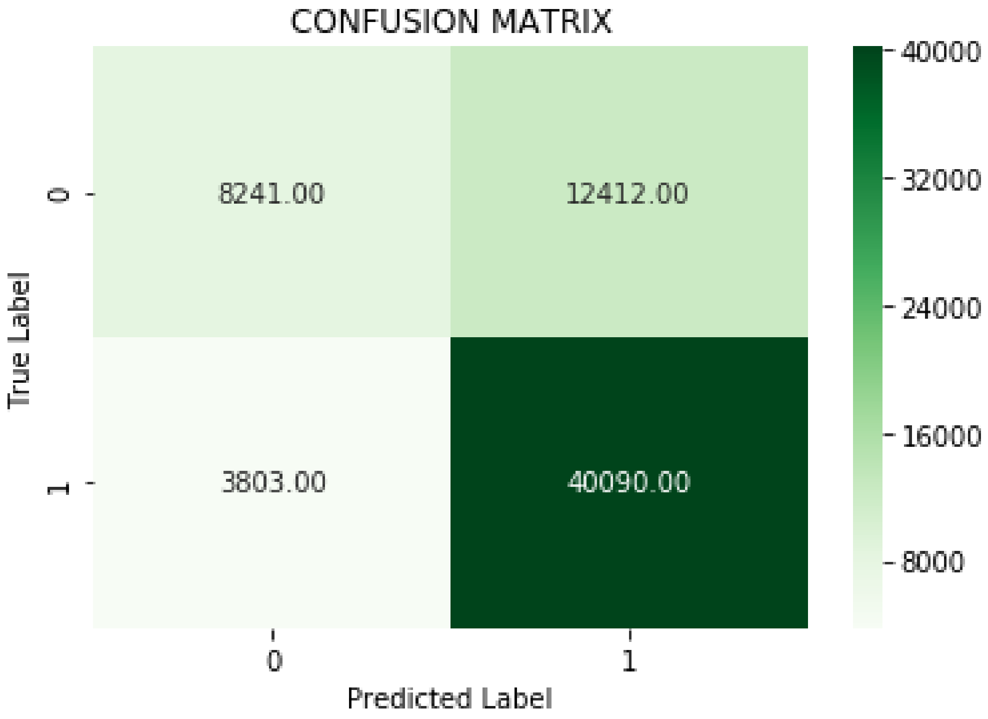
	coef	Odds_ratio
business_stars	1.39	4.0
reviewer_score	0.84	2.3
reviewer_cool	0.00	1.0
business_review_count	0.00	1.0
reviewer_funny	0.00	1.0
reviewer_useful	0.00	1.0
reviewer_review_count	0.00	1.0
new_categories_Spa	-0.35	0.7
new_categories_Automotive	-0.38	0.7
new_categories_Other Foods	-0.47	0.6
business_city_Mesa	-0.69	0.5
business_city_Phoenix	-0.73	0.5
new_categories_Travel	-0.75	0.5
business_city_Scottsdale	-0.75	0.5
business_city_Glendale	-0.75	0.5
business_city_Chandler	-0.76	0.5
business_city_Tempe	-0.83	0.4
new_categories_Bars	-0.97	0.4
new_categories_Restaurant	-0.99	0.4
new_categories_Service	-0.99	0.4
new_categories_Shopping	-1.21	0.3
const	-6.45	0.0

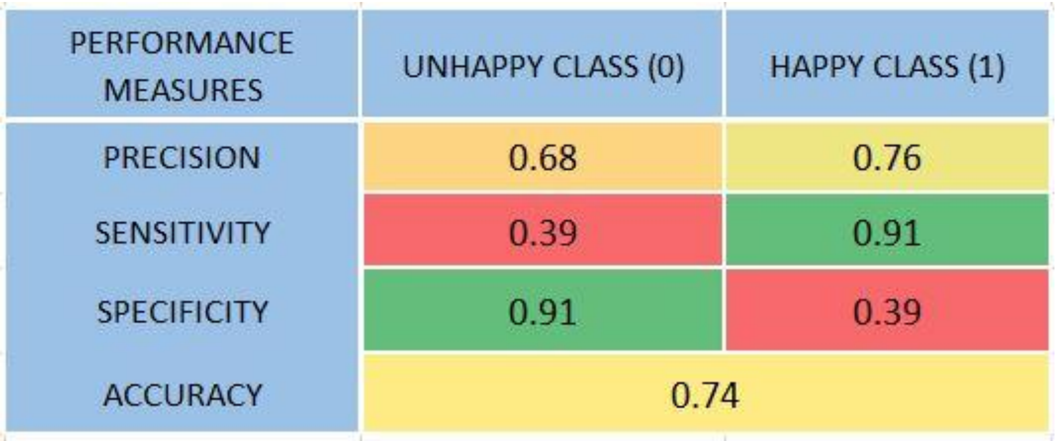
Odds are determined from probabilities and range between 0 and infinity. Odds are defined as the ratio of the probability of success and the probability of failure. The odds of success are

odds(success) = p/(1-p)

Odds are calculated as $B_0 + B_1*x_1 + B_2*x_2$ + ….,which means the odds are highly dependent on business_stars, new_categories, etc.

Confusion Matrix and Performance Measures for Text Classification Model

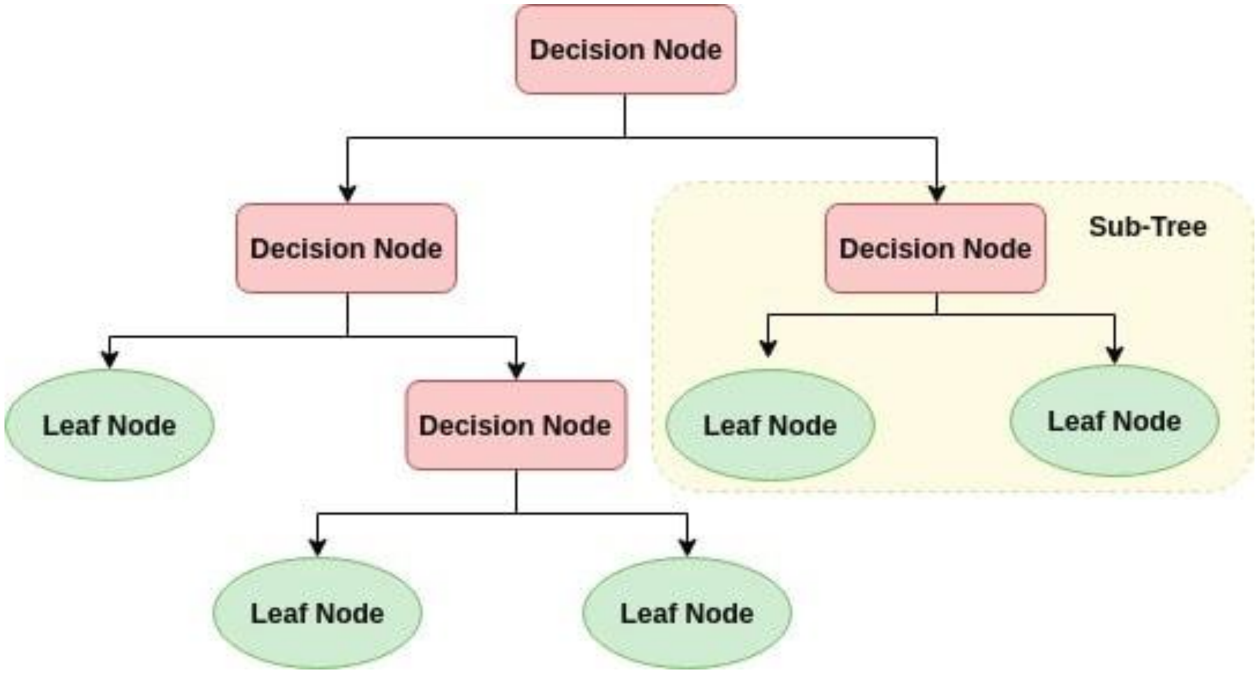




PERFORMANCE MEASURES	UNHAPPY CLASS (0)	HAPPY CLASS (1)
PRECISION	0.68	0.76
SENSITIVITY	0.39	0.91
SPECIFICITY	0.91	0.39
ACCURACY	0.74	

CHAPTER- 4 DECISION TREE

A decision tree is a flowchart-like tree structure where an internal node represents feature(or attribute), the branch represents a decision rule, and each leaf node represents the outcome. The topmost node in a decision tree is known as the root node. It learns to partition on the basis of the attribute value. It partitions the tree in recursively manner call recursive partitioning. This flowchart-like structure helps you in decision making. It's visualization like a flowchart diagram which easily mimics the human level thinking. That is why decision trees are easy to understand and interpret.

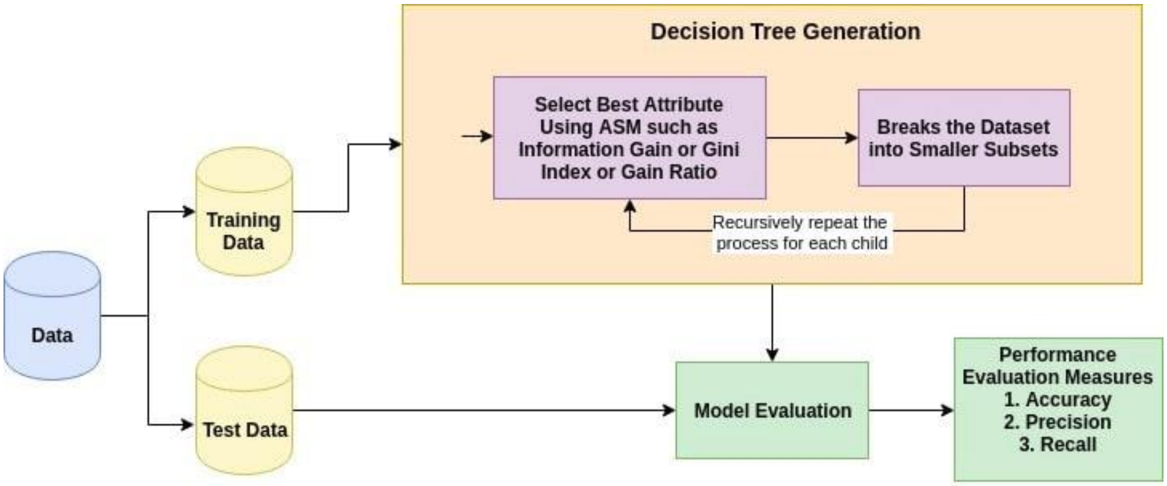


Decision Tree is a white box type of ML algorithm. It shares internal decision-making logic, which is not available in the black box type of algorithms such as Neural Network. Its training time is faster compared to the neural network algorithm. The time complexity of decision trees is a function of the number of records and number of attributes in the given data. The decision tree is a distribution-free or non-parametric method, which does not depend upon probability distribution assumptions. Decision trees can handle high dimensional data with good accuracy.

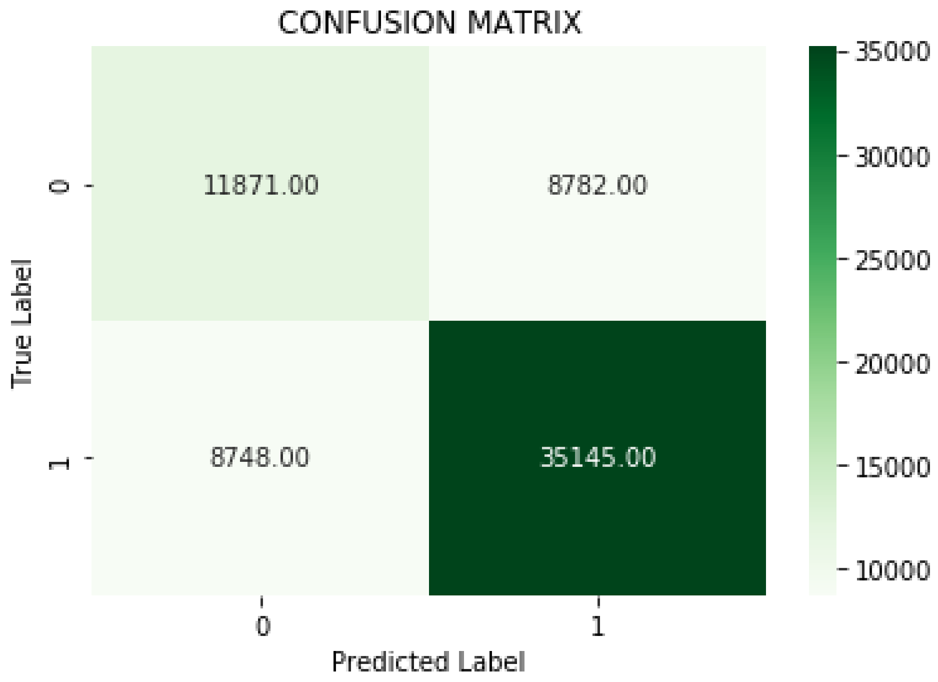
Working of a Decision Tree

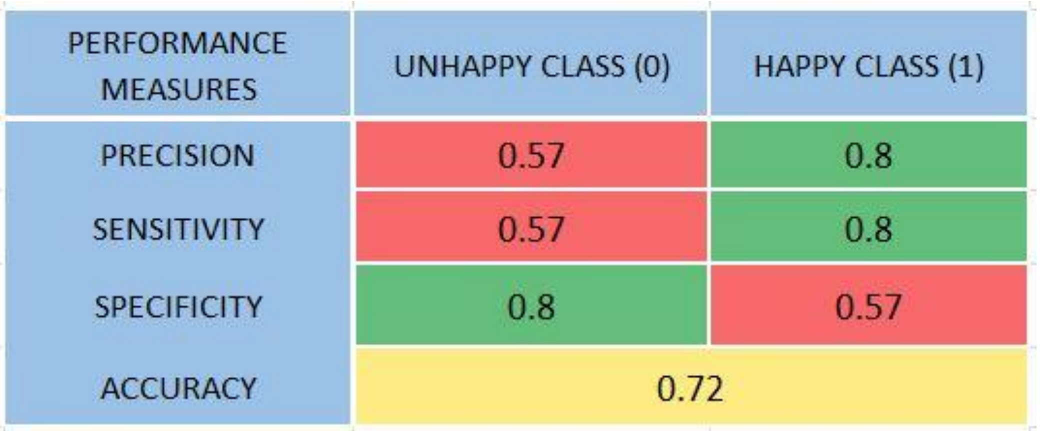
The basic idea behind any decision tree algorithm is as follows:

1. Select the best attribute using Attribute Selection Measures (ASM) to split the records.
2. Make that attribute a decision node and breaks the dataset into smaller subsets.
3. Starts tree building by repeating this process recursively for each child until one of the conditions will match:
   - ✓ All the tuples belong to the same attribute value.
   - ✓ There are no more remaining attributes.
   - ✓ There are no more instances.


Decision Tree Generation
Select Best Attribute Using ASM such as Information Gain or Gini Index or Gain Ratio
Breaks the Dataset into Smaller Subsets
Recursively repeat the process for each child
Data
Training Data
Test Data
Model Evaluation
Performance Evaluation Measures
1. Accuracy
2. Precision
3. Recall

Confusion Matrix and Performance Measures





PERFORMANCE MEASURES	UNHAPPY CLASS (0)	HAPPY CLASS (1)
PRECISION	0.57	0.8
SENSITIVITY	0.57	0.8
SPECIFICITY	0.8	0.57
ACCURACY	0.72	

CHAPTER- 5 RANDOM FOREST

Random Forest is a supervised learning algorithm. It can be used both for classification and regression. It is also the most flexible and easy to use algorithm. A forest is comprised of trees. It is said that the more trees it has, the more robust a forest is. Random forests create decision trees on randomly selected data samples, gets prediction from each tree and selects the best solution by means of voting. It also provides a pretty good indicator of the feature importance.

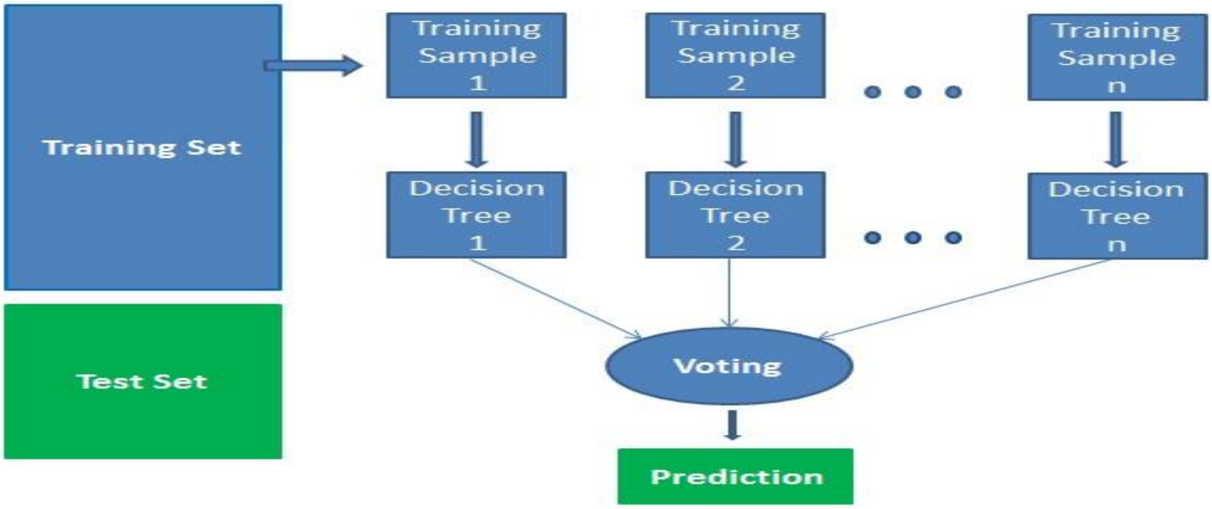
Random forests have a variety of applications, such as recommendation engines, image classification and feature selection. It can be used to classify loyal loan applicants, identify fraudulent activity and predict diseases. It lies at the base of the Boruta algorithm, which selects important features in a dataset.

It technically is an ensemble method (based on the divide-and-conquer approach) of decision trees generated on a randomly split dataset. This collection of decision tree classifiers is also known as the forest. The individual decision trees are generated using an attribute selection indicator such as information gain, gain ratio, and Gini index for each attribute. Each tree depends on an independent random sample. In a classification problem, each tree votes and the most popular class is chosen as the final result. In the case of regression, the average of all the tree outputs is considered as the final result. It is simpler and more powerful compared to the other non-linear classification algorithms.

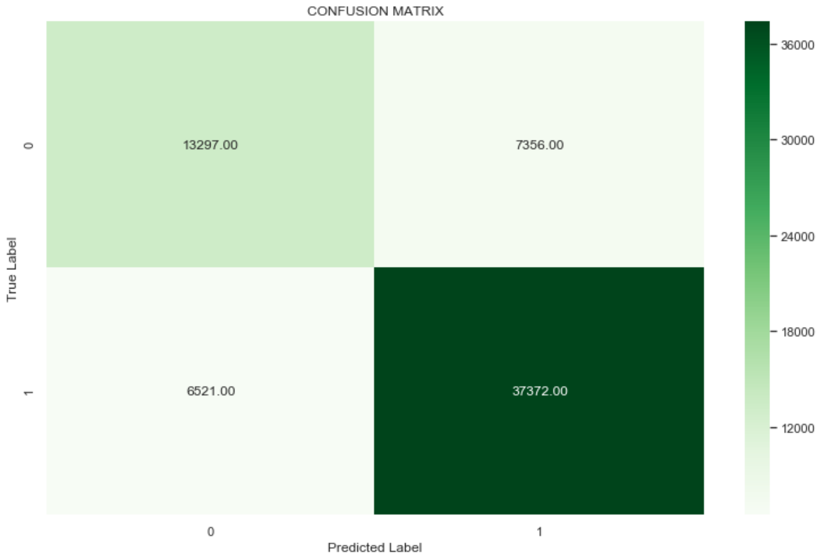
Working of a Random Forest

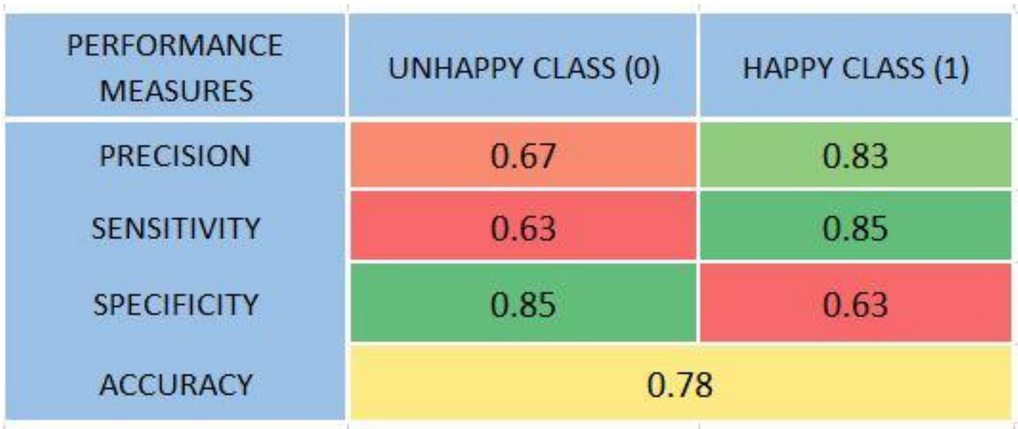
It works in four steps:

1. Select random samples from a given dataset.
2. Construct a decision tree for each sample and get a prediction result from each decision tree.
3. Perform a vote for each predicted result.
4. Select the prediction result with the most votes as the final prediction.

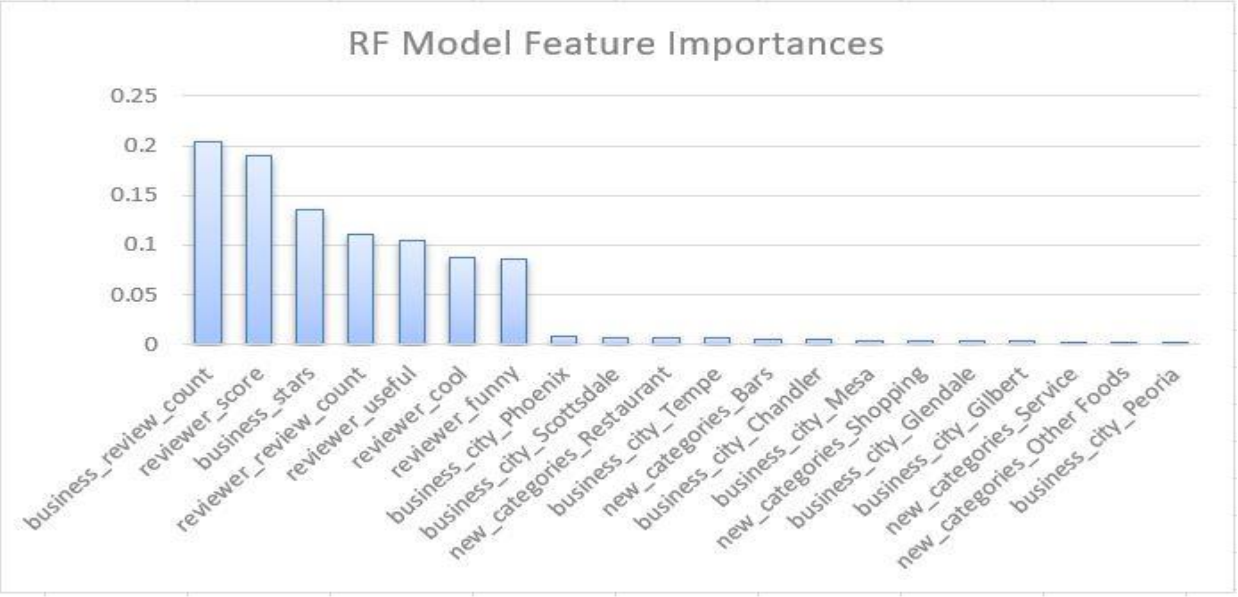


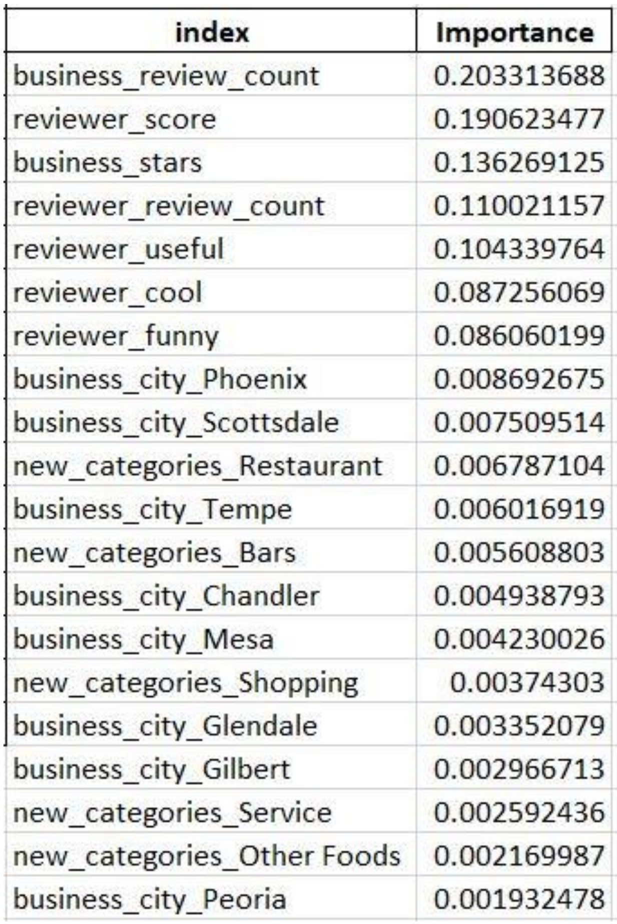
Confusion Matrix and Performance Measures





PERFORMANCE MEASURES	UNHAPPY CLASS (0)	HAPPY CLASS (1)
PRECISION	0.67	0.83
SENSITIVITY	0.63	0.85
SPECIFICITY	0.85	0.63
ACCURACY	0.78	





index	Importance
business_review_count	0.203313688
reviewer_score	0.190623477
business_stars	0.136269125
reviewer_review_count	0.110021157
reviewer_useful	0.104339764
reviewer_cool	0.087256069
reviewer_funny	0.086060199
business_city_Phoenix	0.008692675
business_city_Scottsdale	0.007509514
new_categories_Restaurant	0.006787104
business_city_Tempe	0.006016919
new_categories_Bars	0.005608803
business_city_Chandler	0.004938793
business_city_Mesa	0.004230026
new_categories_Shopping	0.00374303
business_city_Glendale	0.003352079
business_city_Gilbert	0.002966713
new_categories_Service	0.002592436
new_categories_Other Foods	0.002169987
business_city_Peoria	0.001932478

CHAPTER- 6 NAÏVE BAYES

Naive Bayes is a statistical classification technique based on Bayes Theorem. It is one of the simplest supervised learning algorithms. Naive Bayes classifier is the fast, accurate and reliable algorithm. Naive Bayes classifiers have high accuracy and speed on large datasets.

Naive Bayes classifier assumes that the effect of a particular feature in a class is independent of other features. For example, a loan applicant is desirable or not depending on his/her income, previous loan and transaction history, age, and location. Even if these features are interdependent, these features are still considered independently. This assumption simplifies computation, and that's why it is considered as naive. This assumption is called class conditional independence.

Working of Naïve Bayes

Naive Bayes Classifiers rely on the Bayes' Theorem, which is based on conditional probability or in simple terms, the likelihood that an event (A) will happen *given that* another event (B) has already happened. Essentially, the theorem allows a hypothesis to be updated each time new evidence is introduced. The equation below expresses Bayes' Theorem in the language of probability:

Bayes' Theorem is stated as:

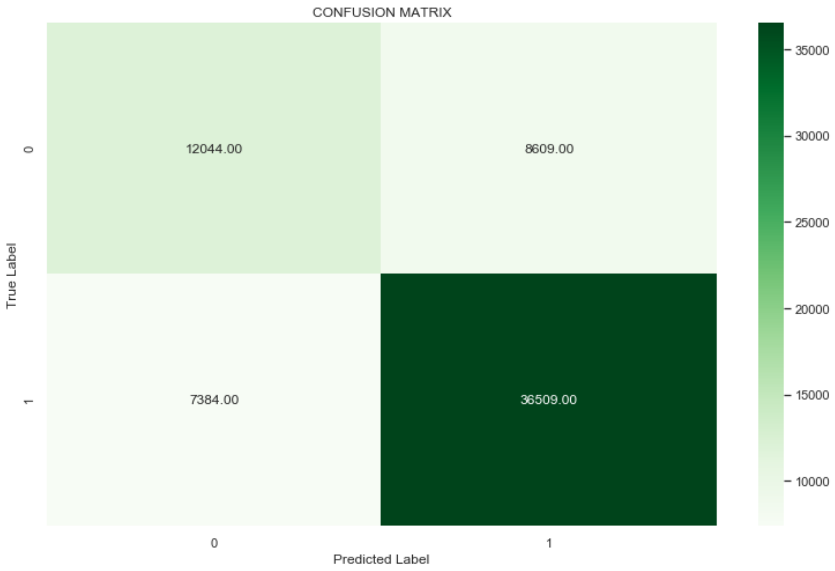
$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

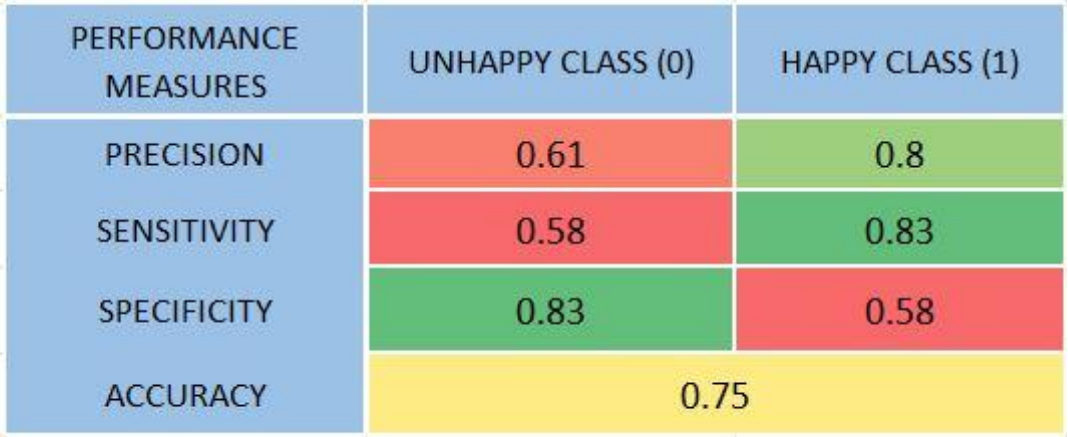
Where

- **P(h/d)** is the probability of hypothesis h given the data d. This is called the posterior probability.
- **P(d/h)** is the probability of data d given that the hypothesis h was true.
- **P(h)** is the probability of hypothesis h being true (regardless of the data). This is called the prior probability of h.

- **P(d)** is the probability of the data (regardless of the hypothesis).

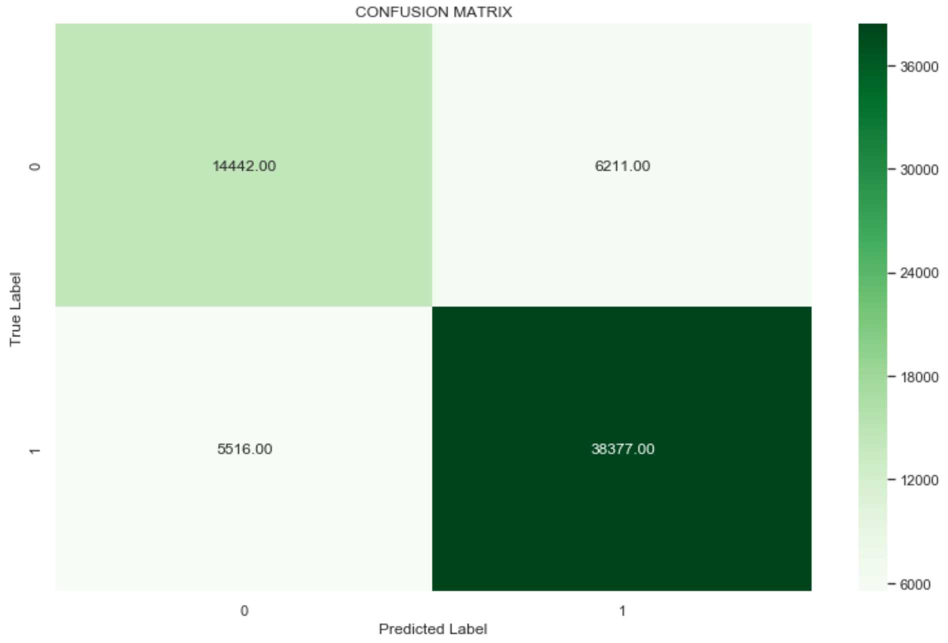
Confusion Matrix and Performance Measures for GaussianNB Classifier

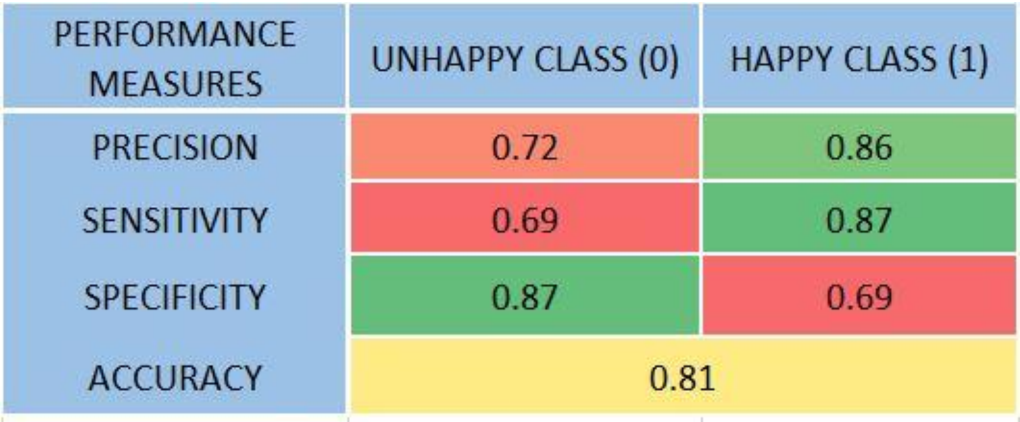




PERFORMANCE MEASURES	UNHAPPY CLASS (0)	HAPPY CLASS (1)
PRECISION	0.61	0.8
SENSITIVITY	0.58	0.83
SPECIFICITY	0.83	0.58
ACCURACY	0.75	

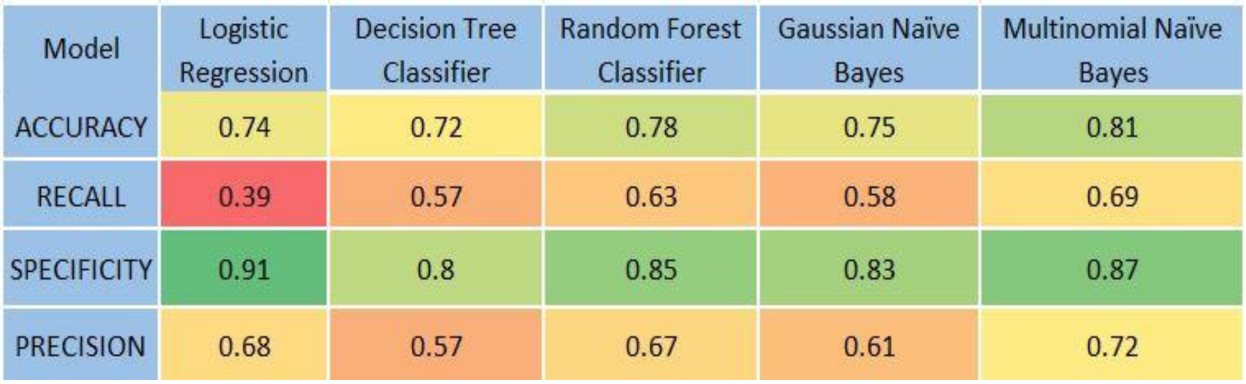
Confusion Matrix and Performance Measures for MultinomialNB Classifier





PERFORMANCE MEASURES	UNHAPPY CLASS (0)	HAPPY CLASS (1)
PRECISION	0.72	0.86
SENSITIVITY	0.69	0.87
SPECIFICITY	0.87	0.69
ACCURACY	0.81	

CHAPTER 7 - COMPARISON OF MODELS



Model	Logistic Regression	Decision Tree Classifier	Random Forest Classifier	Gaussian Naïve Bayes	Multinomial Naïve Bayes
ACCURACY	0.74	0.72	0.78	0.75	0.81
RECALL	0.39	0.57	0.63	0.58	0.69
SPECIFICITY	0.91	0.8	0.85	0.83	0.87
PRECISION	0.68	0.57	0.67	0.61	0.72

Our objective here in our project is to select the Supervised Model which gives us the highest Specificity value for class 0 (Unhappy Customers). This is because our primary motive is to identify the factors which leads to the dissatisfaction of the customers so that we can work on such factors and make the customers satisfied. The reason behind focusing on Specificity over the other performance measures is to reduce the number of FP (false positives) and to restrict us from committing a Type II error which may lead us to lose our valuable customers.

As the formula for Specificity goes:

$$\text{Specificity} = \frac{TN}{TN + FP}$$

The value for specificity will increase when the number of FP's are less. So here below we compare our model parameters for the various Supervised Models we have used for class 0.

Final recommendation on model selection

Logistic Regression is giving us the highest value of Specificity of 91% for class 0. Which is why we go for Logistic Regression as our final model.

CHAPTER 8 - KEY FINDINGS & ACTIONABLE INSIGHTS

- ✓ Happy hours and Friendly customers are highly contributing towards customer satisfaction.

- ✓ Key word like Love, great or good are not providing any meaningful insight. Which we initially thought would provide meaning in deciding if a customer is satisfied.

- ✓ The reviews given by users from Stansfield and Tucson are more trustworthy as they occupy the top 2 spots in terms of reviewer quality.

- ✓ The city of Phoenix has highest number of businesses.

- ✓ The count of business reviews from 2005 to 2012 has increased drastically to more than 60000, but on the contrary, the average ratings have decreased with time from 2005 to 2012.

- ✓ From 2005 to 2006, average ratings have had a sudden drop. After 2006 it is overall maintaining a more or less constant average ratings.

CHAPTER 9 -BIBLIOGRAPHY

Books

- ✓ Business Statistics: A First Course Seventh Edition, Kindle Edition
- ✓ Text Analytics with Python: A Practical Real-World Approach to Gaining Actionable Insights from Your Data

Blogs

- ✓ Medium
    - https://medium.com/analytics-vidhya/automated-keyword-extraction-from-articles-using-nlp-bfd864f41b34
    - https://medium.com/analytics-vidhya/customer-review-analytics-using-text-mining-cd1e17d6ee4e

Websites

- ✓ Data World
  https://data.world/brianray/yelp-reviews/workspace/
- ✓ Yelp
  yelp.com

APPENDIX

For the codes and ipynb python notebook, kindly refer to the github repository below:

https://github.com/Deba04/CAPSTONE-PROJECT