MEF UNIVERSITY

PREDICTING CUSTOMER SATISFACTION VIA STRUCTURED AND UNSTRUCTURED DATA USING CLASSIFICATION AND REGRESSION

Capstone Project

Efehan Danışman

İSTANBUL, 2019

MEF UNIVERSITY

PREDICTING CUSTOMER SATISFACTION VIA STRUCTURED AND UNSTRUCTURED DATA USING CLASSIFICATION AND REGRESSION

Capstone Project

Efehan Danışman

Advisor: Dr. Hande Küçükaydın

İSTANBUL, 2019

MEF UNIVERSITY

Name of the project: Predicting Customer Satisfaction Via Structured And Unstructured Data Using Classification And Regression

Name/Last Name of the Student: Efehan Danışman
Date of Thesis Defense: 13/08/2019

I hereby state that the graduation project prepared by Efehan Danışman (Title Format) has been completed under my supervision. I accept this work as a "Graduation Project".

dd/mm/yyyy
Advisor's Name (Asst. Prof. ...)

I hereby state that I have examined this graduation project by Efehan Danışman which is accepted by his supervisor. This work is acceptable as a graduation project and the student is eligible to take the graduation project examination.

dd/mm/yyyy

Director
of
Big Data Analytics Program

We hereby state that we have held the graduation examination of _____ and agree that the student has satisfied all requirements.

THE EXAMINATION COMMITTEE

Committee Member	Signature
1. Dr. Hande Küçükaydın	……..
2. ……..	……..

Academic Honesty Pledge

I promise not to collaborate with anyone, not to seek or accept any outside help, and not to give any help to others.

I understand that all resources in print or on the web must be explicitly cited.

In keeping with MEF University's ideals, I pledge that this work is my own and that I have neither given nor received inappropriate assistance in preparing it.

Name	Date	Signature
Efehan Danışman	13.08.2019	

EXECUTIVE SUMMARY

PREDICTING CUSTOMER SATISFACTION VIA STRUCTURED AND UNSTRUCTURED DATA USING CLASSIFICATION AND REGRESSION

Efehan Danışman

Advisor: Dr. Hande Küçükaydın

AUGUST, 2019, 24 pages

According to different studies, retaining existing customers is five or more times more costly than acquiring new ones. This study aim to understand what customers expect from an airline using machine techniques. Dataset is scraped from Skytrax's Airline Quality website and consists of 65947 observations with 17 columns consisting of one free format column that includes customer review. In order to do predict whether a customer recommends an airline or not, we try to utilize classification and regression algorithms. In addition to insights, this study also aims to compare the performance of the models and viability of using only free text in order to predict customer satisfaction.

Key Words: Skytrax, airlines, customer satisfaction, classification, regression, machine learning.

ÖZET

MÜŞTERİ MEMNUNİYETİNİ SINIFLANDIRMA VE REGRESYON ALGORİTMALARINI KULLANARAK YAPILANDIRILMIŞ VE YAPILANDIRILMAMIŞ VERİLERLE TAHMİN ETMEK

Efehan Danışman

Tez Danışmanı: Dr. Hande Küçükaydın

AĞUSTOS, 2019, 24 sayfa

Farklı araştırmalara göre firmaların hali hazırdaki müşterilerini elinde tutması yeni müşteri kazanımına göre ortalama beş kat daha maliyetlidir. Bu çalışma makine öğrenmesi yoluyla ve kullanıcıların seçtiği alanlarla serbest metinleri kullanarak müşterilerin havayollarından beklentisini anlamayı amaçlamaktadır. Skytrax'in Airline Quality internet sitesinden alınan veri seti 65947 satır ve 17 sütuna sahiptir. Kullanıcıların bir havayolunu tavsiye edip etmediğini tahmin edebilmek için sınıflandırma ve regresyon algoritmaları kullanılmıştır. Yönetimsel bir kavrayış vermenin yanı sıra çalışma ayrıca farklı algoritmaların performansını karşılaştırmakta ve müşteri memnuniyetini tahmin etmek için serbest metin formatlarının uygunluğunu tartışmaktadır.

Anahtar Kelimeler: Skytrax, havayolları, müşteri memnuniyet, sınıflandırma, regresyon, makine öğrenmesi

Table of Contents

1. INTRODUCTION

1.1. Outline

According to the Pfeier (2005), acquiring new customers costs five times or more than retaining the existing ones. With the exponential development in computation power, gathering of data and the potential of machine learning in extracting insights from a large volume of customer feedback has become much more crucial. For this reason, we aim to extract insights by using a large dataset from Skytrax Airline Quality website (with more than 65.000 observations), where customers leave their impressions regarding an airline they have flown. By extracting insights from data, airlines can understand customer's expectations better and retain them in the long term. Moreover, this work also aims to compare free text features performance to predict whether a customer recommends an airline or not vis-à-vis multiple-choice questions.

This study is organized as follows: Introduction section includes descriptive work about the dataset, shortly presents similar works that have been done as a literature review and explains the objectives of this study. The methodology section describes what has been done with data as pre-processing and feature engineering in order to model the data. Results section discloses advantages and disadvantages of the models with their accuracy and training times. Lastly, the outcomes of the study are presented and recommendations for future works are given.

1.2. Dataset

There are 17 columns and 65947 observations in the dataset. 57 major airlines each with more than 200 observations are selected as target group. Main variables that are used airline company, customer review (free text), review date (date), cabin type (categorical), traveller type (categorical), cabin (categorical), route (categorical), seat comfort (scale 1-5), cabin service (scale 1-5), food and beverage (scale 1-5), entertainment (scale 1-5), ground service (scale 1-5), value for money (scale 1-5) and recommended (yes-no, binary). Data is scraped from airlinequality.com. Example entries can be seen from here.

The dataset has different types of features and some of them have various empty values. Among 17, only 4 of the features do not have any null values. Since not all fields

are obligatory in the review form, null values are natural. In addition, at some airlines, in particular the low-cost ones, entertainment and food and beverage fields do not exist. Table below shows some basic statistics regarding numeric variables of the dataset. From the Table 1, it can be seen that the standard deviation and the means are similar, even though cabin service has slightly higher point than others.

Table 1: Descriptive Statistics of the Numeric Features

	overall	seat_comfort	cabin_service	food_bev	entertainment	ground_service	value_for_money	recommended
mean	5.15	2.95	3.19	2.91	2.86	2.69	2.94	0.49
std	3.48	1.44	1.57	1.48	1.51	1.61	1.59	0.5
min	1	1	1	1	1	1	1	0
0,25	1	1	2	1	1	1	1	0
0,5	5	3	3	3	3	3	3	0
0,75	9	4	5	4	4	4	4	1
max	10	5	5	5	5	5	5	1

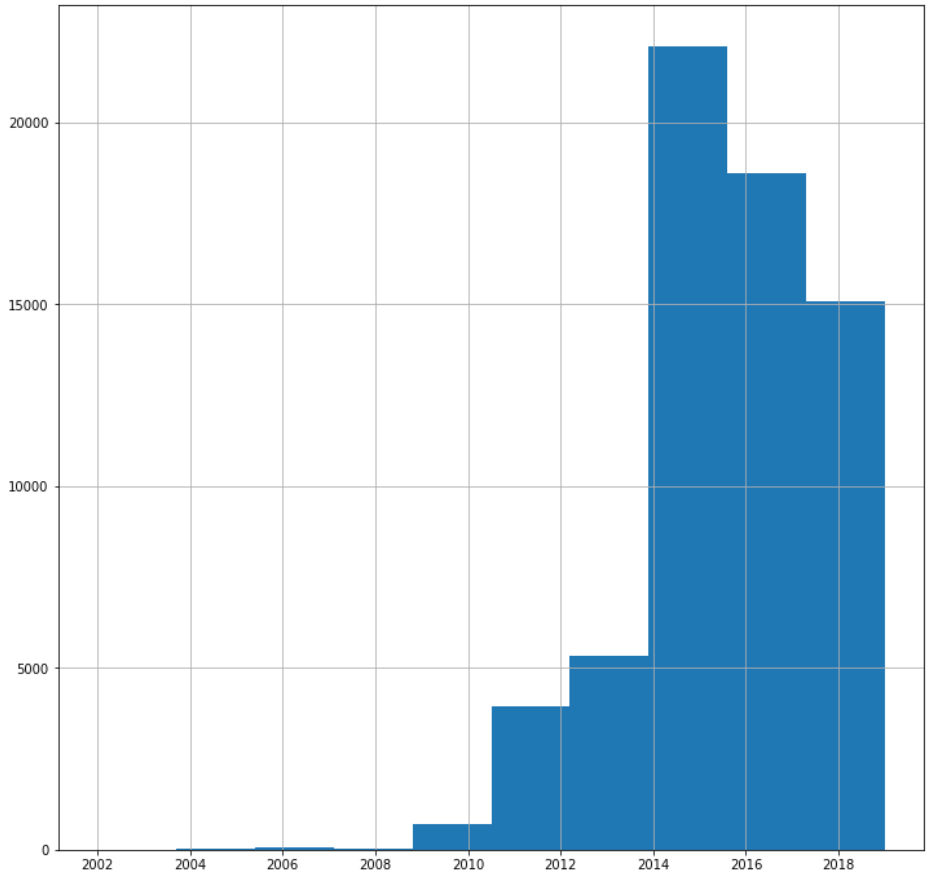
Table 2 presents the number of null values for each feature of the dataset. As you can see except airline, author, review date and customer review variables, every other has at least a thousand missing values. The issue of large missing values is addressed in next section.

Table 2: Number of Missing Values for Each Feature

Feature	Null Values
airline	0
overall	1290
author	0
review_date	0
customer_review	0
aircraft	35294
traveller_type	21142
cabin	1526
route	21165
date_flown	21243
seat_comfort	3553
cabin_service	3529
food_bev	8097
entertainment	14743
ground_service	21429
value_for_money	1497
recommended	1240

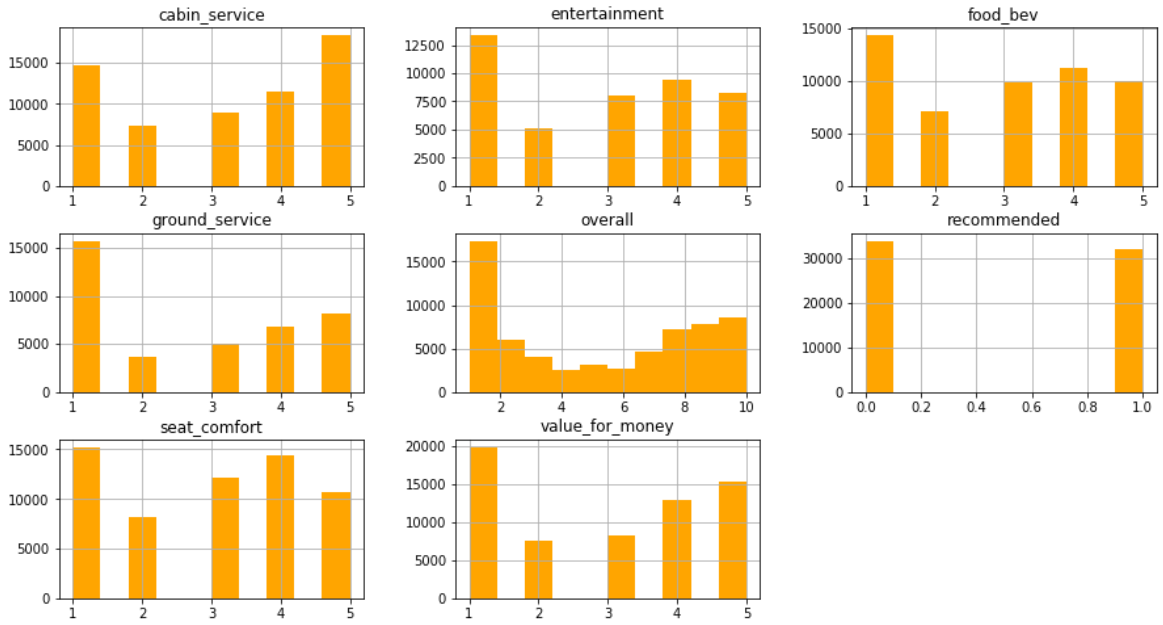
Customer reviews' dates are ranged from 2005 to 2019. However, a significant number of reviews (at least 1000) starts from 2011 as can be seen from the Figure 1. Review data may tell us how customers preferences are evolved from early 2010s till today.

Figure 1: Distribution of the Reviews Across Years



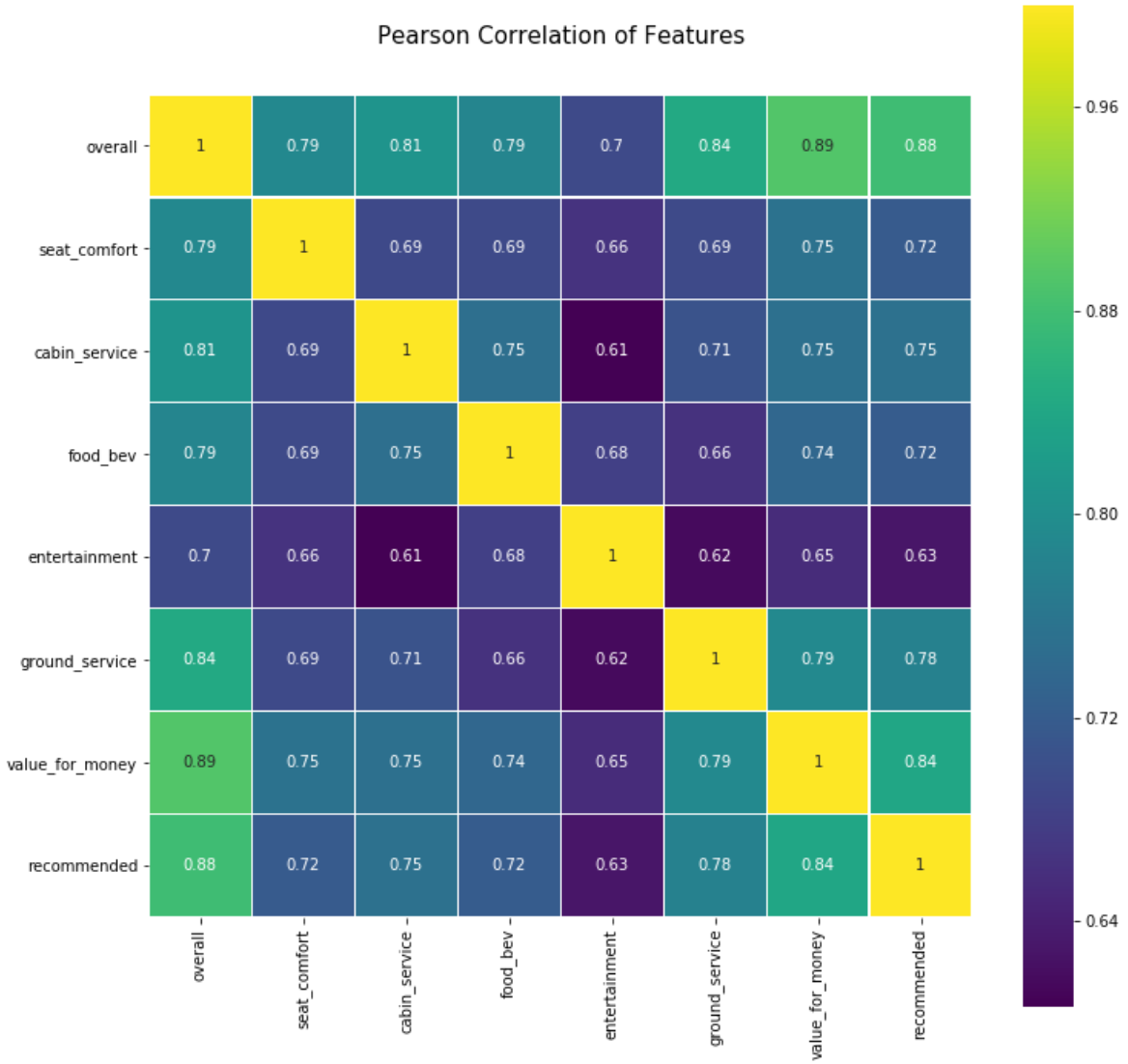
Among airlines, we have seen that customer entries are mostly written for US-based airlines which is in line with its share in commercial aviation. Figure 2 depicts, distributions of the numeric features in 1-5 or 1-10 scale and except for some of the features, most of them are U shaped with highly extreme values along with a few medium-level values. Such a distribution may give us more information when we compare them with our recommended target feature. However, their pairwise relationship are also important. From Figure 2, we can see that our target feature's, which is entitled as recommended, distribution is balanced. That will make our work a bit easier when it comes to preparing our data for a model since there is no need for over or under-sampling.

Figure 2: Distribution of the Numeric Features and Target Feature



After examining the distributions, we concentrate on pairwise correlations between these variables with the help of Figure 3. It looks like all variables are correlated with the target variable recommended to some extent which indicates that using feature selection or dimensionality reduction techniques could be useful to predict the target variable. Among them, the least correlated one is entertainment with 0.63, while the value for money (except overall which is naturally very highly correlated) has the highest correlation with a value of 0.84. Highly correlated predictor variables can be a problem for modelling as a rule of thumb. However, in our case, it does not create a major problem in creating models.

Figure 3: Correlation Between Numeric Features



When we consider the cabin type, we see that most of the passengers, as expected, are economy class passengers. There are also less frequent types of premium cabins, Premium Economy and First Class, all of which can be merged into Business class. Figure 4 demonstrates the distribution of cabin of the passengers. While almost 80% of the passenger reviews belong to the Economy cabin passengers, 20% of them are written by Business class passengers and above.

Figure 4: Distribution of the Flight Cabin

1.3. Objectives

The project has two main objectives. The first one is to find out how we can predict better whether a customer recommends the airline or not via text features vs. categorical and numeric features. In order to test this, first, we use supervised learning algorithms for classification such as Support Vector Machines, Naïve Bayes, Random Forest, and Decision Trees. Target features are recommended and overall variables for classification and regression tasks, respectively. By doing this, we would like to predict whether a customer recommends an airline or not without having the need to ask the question explicitly and also find out which aspect of the service has the highest importance such as cabin service, seat comfort, ground service, etc.

After classification with the target variable recommended, our second objective is using regression algorithms (linear, ridge, lasso, decision tree regressor, support vector regressor, etc.) in order to predict Overall feature as a target variable that is graded by travellers in 1-10 scale.

In addition to predicting whether a customer recommends an airline or not, we take variable importance that affects a passenger's preference into consideration.

1.4. Literature Review

There are various works that use customer feedback and reviews in order to predict customer satisfaction and determine which feature affect it mostly. Lacic et al. (2016) used a fraction of the same dataset and applied Naïve Bayes, Random Forest, CART and Hoeffding Tree algorithms. Yakut et al. (2015) also used a smaller sample of the same dataset to cluster reviews and make a regression analysis by taking into account cabin passengers (Economy, Premium Economy, Business, First). According to both analysis, value for money and cabin staff service are the two leading indicators of customer satisfaction. Punel et al. (2019) used the same dataset for dates between 2011 and 2018 (with 40.000 observations) to make an analysis on variations in passengers' expectations of service quality by different regions. In addition to geographical differences between customers, they suggest that economy class passengers give importance to the value for money feature while seat comfort and cabin service are important features for business class passengers.

There are other studies that deal with customer reviews. Sezgen et al. (2019) used the data taken from the Trip Advisor website and used Latent Semantic Analysis to determine factors that affect customer satisfaction in different cabins and airlines. The study argues that depending on cabin class and airline type (full service / low-cost), there are slight variations in expectations. While Business class and full-service airline passengers expect friendly staff behaviour, Economy class passengers prioritize value for money. Nevertheless, they do not find this difference in expectations between full-service and low-cost airlines huge.

Askalidis and Malthouse (2016) suggest that users who are prompted to write an e-mail give up to 0.5-star higher ratings than customers who are self-motivated to write a review. Hence, it can be useful to acknowledge that Skytrax dataset may have a bias towards negativity since all reviews are written by self-motivated customers.

2. METHODOLOGY

To predict whether a customer recommends an airline or not, two base models are developed with different algorithms for classification and regression tasks:

- A model with numeric features without free format
- A model with only features extracted from the free text (sentiment, length, number of adjectives used, count of words, etc.)

To this end, we make use of the variables review date, cabin, seat comfort, cabin service, food and beverages, entertainment, ground service and value for money in the training set. Since not all flights have entertainment and food as a service, we develop our model only for full-service flights with structured data.

Models mentioned above are 10-fold cross-validated and grid search method is implemented if available for the algorithm.

Since there are many null values, some feature engineering is applied to our features. Empty fields are either imputed or removed (in case this feature is not available for the related flight). For instance, ground service is added to the questions in 2014, since all values are empty before, while cabin service was available.

Feature selection methods are also be applied to models in order to achieve the most accurate and simplest possible model at the same time.

2.1. Feature Engineering, Extraction, Selection

2.1.1. The model with numeric features without free text

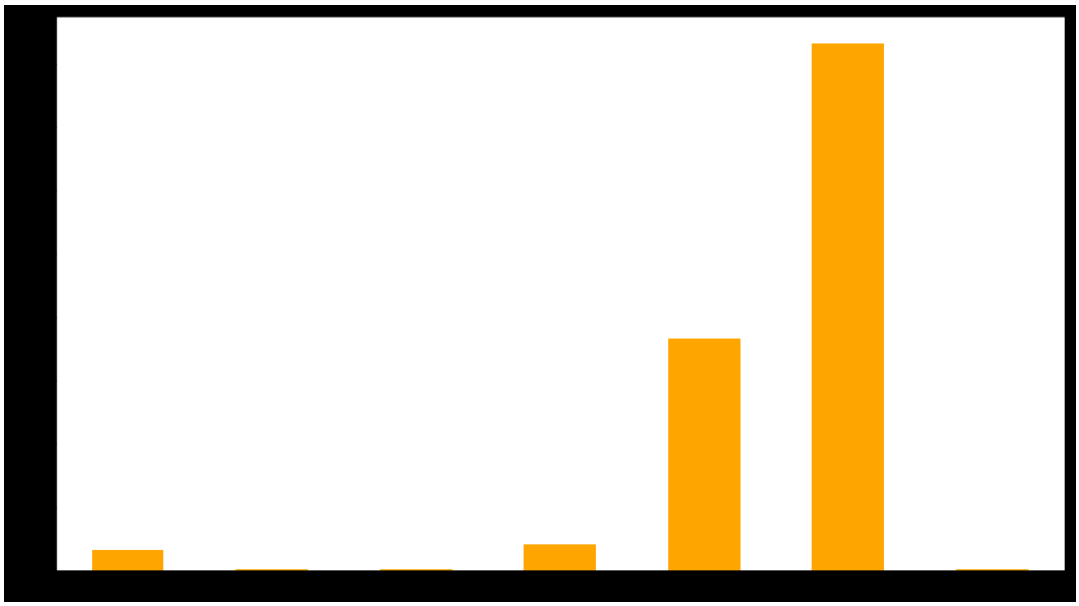
In this part, first, we deal with missing values. As a starting point, we see that rows with 60% of missing values, do not have any useful information for our models. Hence, they are discarded. As a result, the number of observations is decreased from 65947 to 60444.

Secondly, food and beverage along with entertainment are not available for each observation since not all flights have such services. In order to model full-service flights, where entertainment and food beverage are offered to passengers, we removed

observations where both fields are missing. After this operation, the total number of observations decreased to 53,188.

Even after these operations, we are left with a considerable amount of missing values for each column which can be observed from Figure 5 beluw. Each of the features has some missing values, while ground services and entertainment lead with a large margin.

Figure 5: Distribution of the Missing Values



In order to deal with missing values, we used MissForest algorithm developed by Stekhoven and Bühlmann (2012) which fills missing values using Random Forest algorithm in order to predict them based on out-of-bag (OOB) imputation error. In this context, main advantages of using MissForest as an imputer can be explained follows. It is a non-parametric model which does not assume that distributions are normal and it is not restricted to a single variable as opposed to other imputers.

For categorical features we have, the cabin is merged into Economy and Business while 890 observations (around 2% of all observations), kept as Unknown variable.

For traveller type, Business and Leisure types are kept and 20.000 missing values which is almost 40% of the whole dataset. As in Cabin variable, here I also added an Unknown feature since the effect of the variables is not significant vis-à-vis target value.

In order to transform the data into a modellable format, we one hot encoded Cabin and Traveller Type variables. Airline variable is also dropped due to the number of levels since one hot encoding with many different levels brings us to the curse of dimensionality.

2.1.2. The model with only free text features

Predicting whether a customer recommends an airline or not is significantly harder with solely using free format features. For this reason, the text is cleaned up and various feature extraction and dimensionality reduction techniques are employed.

Firstly, word and character length of the customer review have been extracted as a feature. The text is lowercased, numbers and symbols (e.g. currency signs) are removed along with certain stop words.

Next, by using nltk package, words are lemmatized. Then Part of Speech (POS) tags are used to extract various features in form of count such as NN (noun), POS (possessive endings), JJ (adjectives), JJS (comparatives), etc. 28 different features are extracted in this way. List of the POS tags can be found from the University of Penn's website[1].

After that, count of each word's that exists in at least 2.5% of the texts extracted as the feature as a result, it expanded the dimensionality towards 649 columns with 65947 rows. Such a data would easily fall into the curse of dimensionality – the need for extra observations as the number of features increased exponentially – along with computational complexity. Hence, we decided to implement some dimensionality reduction techniques.

To reduce dimensionality, we use sci-kit learn package's Variance Threshold method. We set minimum variance to 0.15 which filters out all variables below that threshold. This left us with 146 features which means that most of the columns have little or no information.

From here, we create several base models. However, in order to improve computational efficiency, we use the default Random Forest model and get feature importance of the variables from there. For feature importance, we used Mean Decrease in Impurity and Permutation Importance criteria. Mean decrease in impurity is looking for

[1] Retrieved on 09.08.2019, https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

best splits where most of the splits recommend (or not) an airline. On the other hand, permutation importance shuffles the variables and look for a decrease in accuracy when a certain variable is removed from the model.

So, we select all variables which have at least 1% importance and this causes only a 2% decrease in test accuracy with 30% of the dataset which is explained in more detail in the next section. However, the Random Forest model's default feature importance based on the mean decrease in impurity has a certain bias when features are correlated, or the number of categories is different for features. To double-check the feature importance here we also check Permutation Importance (loss of accuracy when the related variable is withdrawn from the model) and their results are similar with the random forest's feature importance. You can see the details of the feature importance with Permutation Importance and Mean Decrease in Impurity criteria by examining Table 3 and Table 4. At the end of the feature selection, we are left with 16 features that are a mix of word counts, certain words and part of speech tags as the best predictors of whether recommending an airline or not.

Table 3: Variable Importance in terms of Permutation Importance

0.1247 ± 0.0012	good
0.0653 ± 0.0016	excellent
0.0559 ± 0.0014	comfortable
0.0532 ± 0.0016	friendly
0.0488 ± 0.0016	great
0.0486 ± 0.0010	hour
0.0478 ± 0.0014	never
0.0464 ± 0.0014	word_count
0.0352 ± 0.0008	told
0.0284 ± 0.0013	crew
0.0280 ± 0.0008	customer
0.0259 ± 0.0009	VBP
0.0226 ± 0.0005	nice
0.0154 ± 0.0009	text_lengthc

Table 4: Variable Importance in terms of Mean Decrease in Gini Impurity

weight	feature
0.116128	good
0.068895	excellent
0.045271	friendly
0.043775	comfortable
0.034395	word_count
0.031966	never
0.030884	great
0.029497	told
0.023214	text_lengthch
0.018684	hour
0.016776	NN
0.015335	VBP
0.014634	customer
0.013787	VBD
0.013449	RB
0.012814	nice
0.011828	VBG
0.010268	crew
0.010180	VB

3. RESULTS AND DISCUSSION

3.1. The model with user selected features without free text-Classification

After reaching the state of the dataset we desired, several models are employed to obtain the best possible result. 10-fold stratified k-fold cross-validation and Grid Search is employed for each dataset. We try to start from the simplest model to the most complex one while making predictions and results are summarized at Table 5.

Table 5: Classification Results with User Selected Features

MODEL	TEST RMSE	TRAINING RMSE	F1-SCORE	BEST PARAMETERS (IF AVAILABLE)
GAUSSIAN NAÏVE BAYES	93.2%	93.2%	93.0%	Not available.
SUPPORT VECTOR MACHINES	94.2%	94.3%	94%	C:1, gamma: 'scale', kernel='rbf'
DECISION TREE	93.7%	94.1%	94%	Max depth:10, min samples leaf: 45
RANDOM FOREST	94%	94%	94%	Max depth:10, min_samples_leaf:100, min_samples_split:20, n_estimators:50
NEURAL NETWORKS	93%	93%	%93	One layer, 150 neurons with softmax function. Learning rate: 0.01

All the models performed well since the prediction is straight forward with a few scaled variables. When model complexity increases, the performance only increases by 1% when compared to the simplest Naïve Bayes model and depending on the importance of this 1%, the time needed to spend can be crucial or a waste of time and computation power. All models including Neural Network's scores differ only in decimal points.

We checked features' importance in the model using Permutation Importance with eli5 package of Python. We do not use Random Forest's default variable importance since when features correlated, it can show important variables less important than they are, as discussed above. Unlike other studies, we have seen that the most important variable is not value for money for Economy passengers. Nonetheless, for business passengers, ground and cabin services are the following most important variables after value for money. Table 6 and Table 7 display variable importance for Economy and Business class passengers. There is not any significant difference except in weights when we do the same analysis for all passengers.

Table 6: Permutation Importance for Economy Class

Weight	Feature- Economy Class
0.0522 ± 0.0038	ground_service
0.0415 ± 0.0016	value_for_money
0.0137 ± 0.0013	cabin_service
0.0068 ± 0.0011	seat_comfort
0.0062 ± 0.0005	food_bev
0.0029 ± 0.0011	entertainment
0.0008 ± 0.0003	type_Solo Leisure
0.0003 ± 0.0004	type_Unknown
0.0002 ± 0.0001	type_Business
0.0001 ± 0.0003	type_Family Leisure
0.0000 ± 0.0003	type_Couple Leisure
0 ± 0.0000	cabin_Unknown
0 ± 0.0000	cabin_Economy Class

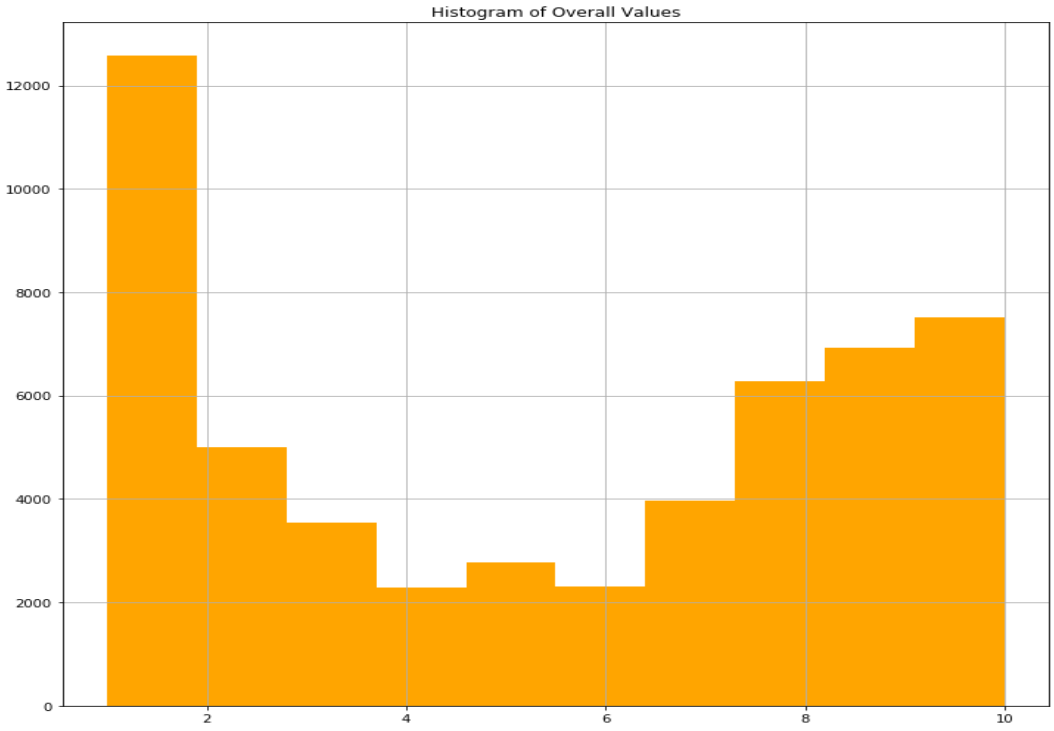
Table 7: Permutation Importance for Business Class

Weight	Feature – Business Class
0.2033 ± 0.0080	value_for_money
0.0128 ± 0.0018	ground_service
0.0087 ± 0.0022	cabin_service
0.0047 ± 0.0030	seat_comfort
0.0040 ± 0.0017	food_bev
0.0004 ± 0.0005	type_Solo Leisure
0.0003 ± 0.0005	type_Couple Leisure
0.0002 ± 0.0004	type_Unknown
0.0002 ± 0.0014	entertainment
0.0001 ± 0.0004	cabin_Business Class
-0.0001 ± 0.0002	cabin_Unknown
-0.0001 ± 0.0001	type_Family Leisure
-0.0003 ± 0.0004	type_Business

3.2. The model with user selected features without free text – Regression

For regression, we do same pre-processing steps as we did for the previous model. However, this time our target feature is switched from recommended to overall. It is the point between 1-10 scale given by passengers to the flight. Figure 6 clearly shows its U-shaped distribution where zeroes are dominant.

Figure 6: Distribution of the Target Value - Overall



In order to test our model, we utilize simple linear, lasso, ridge and random forest regressors. The result is a bit far less accurate from the classification model since regression is a harder problem comparing to classification for our problem. You can see results of the models and parameters we have used at Table 8.

Table 8: Regression Results with User Selected Features

MODEL	TEST MSE	TRAINING RMSE	R^2	BEST PARAMETERS (IF AVAILABLE)
LINEAR REGRESSION	1.14	1.14	0.89	-
LASSO REGRESSION	1.14	1.14	0.90	Alpha= 2.26
RIDGE REGRESSION	1.14	1.13	0.89	Alpha= 0.0005
RANDOM FOREST	1	0.97	0.91	Max_depth=15, min_samples_leaf=50, min_samples_split=100, n_estimators=1000

Since model with user-selected features is rather straightforward, all types of regressions resulted similar as can be seen from Table 8 above even with highly different alpha values. As a result of this alpha, several variables coefficients becomes zero (or close to zero) as can be seen from the formula of the model. According to the regression model, the importance of variables does not differ significantly compared to the classification model we have built previously. Below you can see formulas of the Lasso and Ridge models we have built. Note that variable that are 0 with Lasso model are almost zero in Ridge regression. On the other hand, coefficients are extremely high for one-hot-encoded variables with linear regression even though results are similar.

Linear Regression model: 3563858383892.0 * cabin_Business Class + 3563858383892.0 * cabin_Economy Class + 3563858383892.0 * cabin_Unknown + 2353734412418.0 * type_Business + 2353734412418.0 * type_Couple Leisure + 2353734412418.0 * type_Family Leisure + 2353734412418.0 * type_Solo Leisure + 2353734412418.0 * type_Unknown + 1.0 * ground_service + 1.0 * value_for_money + 0.0 * seat_comfort + 0.0 * cabin_service + 0.0 * food_bev + 0.0 * entertainment

Lasso model: 0.796 * value_for_money + 0.65 * ground_service + 0.353 * cabin_service + 0.248 * seat_comfort + 0.241 * cabin_Unknown + 0.237 * food_bev + 0.114 * entertainment + 0.053 * type_Unknown + -0.033 * cabin_Economy Class + -0.007 * type_Family Leisure + -0.004 * type_Couple Leisure + 0.0 * cabin_Business Class + -0.0 * type_Business + 0.0 * type_Solo Leisure

Ridge model: 0.795 * value_for_money + 0.65 * ground_service + 0.354 * cabin_service + 0.279 * cabin_Unknown + 0.249 * seat_comfort + 0.238 * food_bev + 0.116 * entertainment + 0.043 * type_Unknown + -0.032 * type_Family Leisure + -0.03 * cabin_Economy Class + -0.028 * type_Couple Leisure + -0.02 * type_Business + 0.009 * cabin_Business Class + -0.007 * type_Solo Leisure

Finally, we check the significance of these variables in a linear model and none of the features look insignificant with *p-values* around 0. Table 9 below shows regression coefficients for each variable and their *p-values*.

Table 9: Summary Results of the Linear Regression

REGRESSION COEFFICIENTS						
	coef	std err	t	P>|t|	[0.025	0.975]
const	12.431	0.014	-86.996	0.000	-1.271	-1.215
seat_comfort	0.2529	0.007	35.909	0.000	0.239	0.267
cabin_service	0.3459	0.007	51.697	0.000	0.333	0.359
food_bev	0.2454	0.007	34.275	0.000	0.231	0.259
entertainment	0.1115	0.006	17.862	0.000	0.099	0.124
ground_service	0.6424	0.007	90.213	0.000	0.628	0.656
value_for_money	0.8013	0.008	104.763	0.000	0.786	0.816
cabin_Business Class	0.4826	0.017	-28.913	0.000	0.515	0.45
cabin_Economy Class	0.521	0.014	-36.033	0.000	0.549	0.493
cabin_Unknown	0.2394	0.036	-6.582	0.000	0.311	0.168
type_Business	0.2471	0.016	-15.802	0.000	0.278	0.216
type_Couple Leisure	0.2698	0.014	-19.805	0.000	0.296	0.243
type_Family Leisure	0.2834	0.016	-18.233	0.000	0.314	0.253
type_Solo Leisure	0.2457	0.012	-20.249	0.000	0.27	0.222
type_Unknown	0.1971	0.010	-19.693	0.000	0.217	0.177

3.3. The model with free text – Classification

From the feature engineering section, we have seen that we create a high dimensional dataset from text features. The first model we create is a Random Forest with 146 features and accuracy is 82% on the test set, with 98% on the training set. However, the model is overfitted and it takes time to compute on a high dimensional dataset.

To solve the problems mentioned above and to select the best features, we first select features that create at least 0.01 increase in mean Gini impurity. This decreased the number of features to 16 and a new model created with Random Forest has %79 test accuracy with 84% training accuracy. Before modelling, we used a min-max scaler to scale the values in order to suppress over-dominance of one feature comparing to others. We do not overfit now and gain a significant amount of computation time by losing some accuracy on the test set. It looks plausible to go on with 16 variables rather than 164 if the problem we are dealing with does not extremely sensitive to the small changes such as fraud detection.

According to results with ensemble methods and support vector machine performed well in the test set using only a few variables and hyperparameter tuning. None of the models overfitted and performed satisfactory predictions. Table 10 shows result of these models we have built.

Table 10: Classification Results with Free Text

MODEL	TEST RMSE	TRAINING RMSE	F1-SCORE	BEST PARAMETERS (IF AVAILABLE)
GAUSSIAN NAÏVE BAYES	77%	76%	77%	-
SUPPORT VECTOR MACHINES	79%	80%	79%	C:1, gamma: 0.1, kernel='rbf'
RANDOM FOREST	79%	84%	79%	Max depth:13, min_samples_leaf:50, min_samples_split:10, n_estimators:50
GRADIENT BOOSTING	80%	82%	78%	Learning rate = 0.25, min_samples_split = 500, n_estimators = 500

3.4. The model with free text - Regression

With the 16 features we selected above, we run the same models with selecting Overall (1-10 scale) feature as the target variable. Even before selecting variables after removing variables with low variance, we have 146 variables. Since we already have low performance, we do not go further down to limit the number of variables. Table 11 summarizes the results of several models.

With grid search and 10-fold cross-validation, we slightly improved the performance, but it is still far from the performance of regression with user-selected features. The best result is obtained by the tuned Random Forest model rather than linear models and it is overfitted. Boosting methods could be employed in order to improve these results with careful tuning.

Table 11: Regression Results with Free Text

MODEL	TEST MSE	TRAINING RMSE	R^2	BEST PARAMETERS (IF AVAILABLE)
LINEAR REGRESSION	2.50	2.47	0.48	-
LASSO REGRESSION	2.50	2.47	0.48	Alpha:5.61
RIDGE REGRESSION	2.50	2.47	0.49	Alpha: 0.001
RANDOM FOREST REGRESSION	2.23	1.36	0.59	Max_depth = 30, min_samples_split = 5, n_estimators = 200, min_samples_leaf = 2

4. CONCLUSION AND FUTURE RESEARCH

We have several takeaways from this study. As expected, the first one is, that is much easier to predict whether a customer recommends an airline or not with user-selected features. However, we obtain the most satisfactory results using free text fields as well by predicting over 80% of the test set with only 16 features.

For user-selected fields, we have seen that value for money, ground service and cabin service are the most important fields (with a different order for Economy and Business class passengers). Thus, for customer satisfaction, pricing perception and expectations of the customers has the utmost value. Moreover, service in-ground and in-flight (particularly cabin crew) have high importance. On the other hand, entertainment and catering are also important with relatively less influence. Lastly, cabin type and traveler type have no impact on customer's review comparing to other variables.

From the machine learning perspective, when the problem is straightforward with the data we have, using complex methods such as Neural Networks do not bring any satisfactory improvement. However, there are cases such as dealing with free text fields where using more complex methods could bring meaningful improvement.

On the regression side, we have explained 89% of the variation with the user-selected features and minimized our root mean squared error up to 1 out of 10 using Random Forest Regressors. Generalized linear models (Lasso and Ridge) also reached satisfactory results with 1.05. Like classification models, one-hot-encoded variables such as cabin, traveler type do not have a high impact compared to other variables in our models.

For classification with free text fields, even with only certain words occurrence, word, and character length, we can predict mostly whether a customer recommends an airline or not. With vectorizing words and getting part of speech tags, we reached 80% accuracy. Here we have seen that, customers tend to write longer when they have a negative review. On the regression side, it is more challenging since results were not highly satisfactory. By using extreme gradient boosting or deep learning algorithms, this can easily improve towards similar levels with classification on user-selected features.

With the same dataset, future research can be done on classifying free text fields or improving classification and regression performance without needing user-selected features via using deep learning.

REFERENCES

- Airlinequality website, example entry, https://www.airlinequality.com/airline-reviews/air-france, Accessed on 27th July, 2019.
- Alphabetical list of part-of-speech tags used in the Penn Treebank Project, https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html, Accessed on 9th August, 2019.
- Askalidis, G., Kim, S.J. and Malthouse, E.C. (2017), Understanding and Overcoming Biases in Online Systems, *Decision Support Systems*, Vol.97, 23-30.
- Bird, S., Klein, E. & Loper, E., (2009), Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit, O'Reilly.
- Geron, Aurelie (2017), Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools and Techniques to Build Intelligent Systems, O'Reilly.
- Lacic, E., Kowald, D. & Lex, E. (2016), High Enough? Explaining and Predicting Traveller Satisfaction Using Airline Reviews, *Proceedings of the 27th ACM Conference on Hypertext and Social Media*, 249-254.
- Pfeier, P.E. (2005), Optimal Ratio of Retention and Acquisition Costs, *Journal of Targeting, Measurement and Analysis for Marketing*, Vol. 13, 1479-1862.
- Punel, A., Hassan, L.A.H., Ermagun, A. (2019), Variations in Airline Passenger Expectation of Service Quality Across the Globe, *Tourism Management*, Vol.75, 491-508.
- Sezgen, E., Mason, K.J. & Mayer, R., (2019), Voice of Airline Passenger: A Text Mining Approach to Understand Customer Satisfaction, *Journal of Air Transport Management*, Vol. 77, 65-74.
- Stekhoven, D. & Bühlmann, P. (2012), MissForest – Non-parametric Missing Value Imputation for Mixed-type Data, *Bioinformatics*, Vol.28, 112-118.
- Strobl, C., Boulesteix, A., Zeileis, A. & Hothorn T., (2007), Bias in Random Forest Variable Importance Measure: Illustrations, Sources and a Solution, *BMC Bioinformatics*, 8-25.

- Yakut, I, Turkoglu, T. & Yakut, F. (2015), Understanding Customers' Evaluations Through Mining Airline Reviews, *International Journal of Data Mining & Knowledge Management Process*, Vol.5, No.6, 1-11.

TABLES AND FIGURES

LIST OF TABLES

LIST OF FIGURES