DECEMBER 2019

Predictions for airline reviews

by Monika Krajnc

Springboard Data Science course

Overview

The Main Question

Positive reviews are becoming increasingly important, since a large portion of customers are basing their purchase decision on them.

Understanding customer feedback is thus an important component of better serving them, but offers only a snapshot of the current situation.

Would it be possible for the airline to ***predict future customer reviews based on the current situation****, and better anticipate potential issues?*

The Problem for Airlines

Airlines are trying to offer the best possible experience and comfort without creating additional costs for offering it. Therefore it's important for them to understand the customer satisfaction with the service. And one way to track it is through unsolicited reviews. But how can they better understand if an event was a one-off or is indicative of a systematic problem?

The Problem for Airlines

Airlines need to get a better understanding what their passengers care about the most, right now and in the future. And then focus on the most critical issues that will make the most difference, to make flying less of an ordeal and consequently charge more for a better service.

Benefits for Passengers

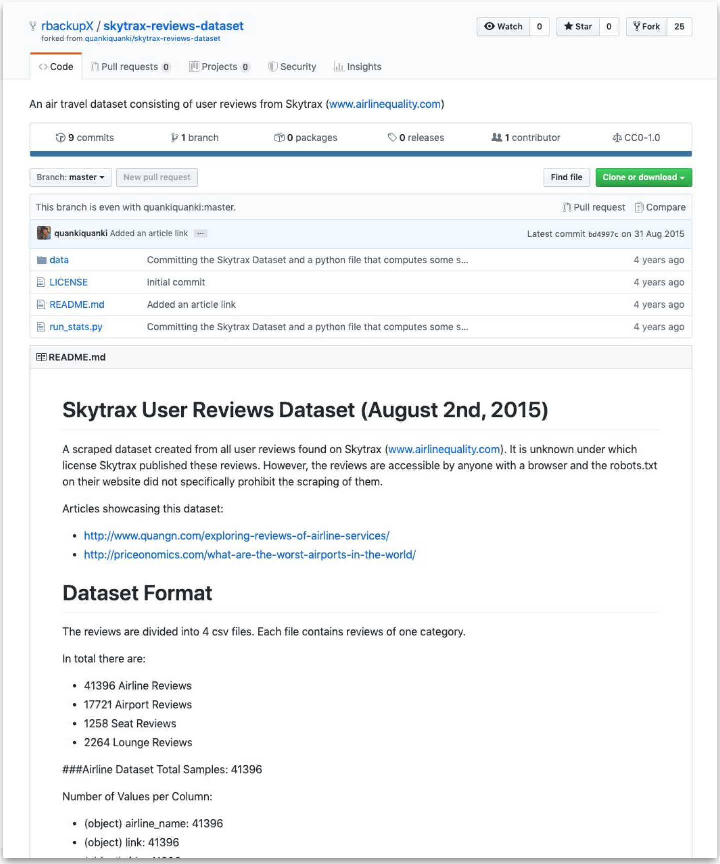
Passengers need help deciding how to pick the right airline, one that has positive reviews for the things they care about most. One interesting aspect would be if the system could generate a review the passenger would write before even taking the flight, based on their past reviews as well as reviews of other passengers. This would also encourage passengers to write honest reviews.

Benefits for Staff and Crew

Not all initiatives need to come from the top, and giving the flight crew better insights and tools is a good place to start. Providing them with insights into which part of the service needs improvement, and giving them the tools to understand the effect of current actions on future outcomes through predicted reviews, would be beneficial and empowering.

The Data

We based our models on a Skytrax Reviews dataset we discovered on Github. It already contains everything we need from Airline name, Country of passenger, Written review content, Cabin class, Various ratings, as well as if the passenger would recommend this Airline to others.



Exploratory Analysis and Findings

Data Preparation

The original dataset contained reviews for 362 airlines, with 41396 entries and 20 columns – some of which are integers for reviewing different aspects of the services, some traveller information, and one plain text field.

We decided to only use airlines that have at least 100 reviews to ensure that there is enough data to draw any conclusions, which in our case is 112 airlines.

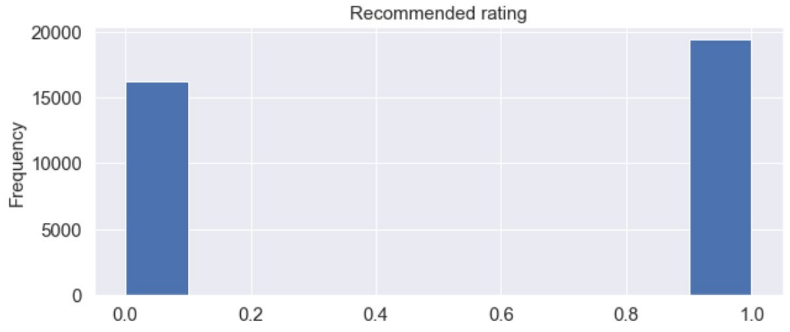
After filtering out the airlines with fewer than 100 review, we are left with 35609 rows and 20 columns.

Our Questions

1. *How does the traveller type (leisure solo, couple, family, or business), and cabin flown affect the rating? Can we notice any patterns?*
2. *Are ratings where travellers recommend an airline connected to overall ratings?*
3. *Are there any differences between traveller type categories and cabin flown categories?*
4. *Does the traveller's country of origin influence the rating?*
5. *Is there a connection between value_money rating and the overall or recommend rating?*

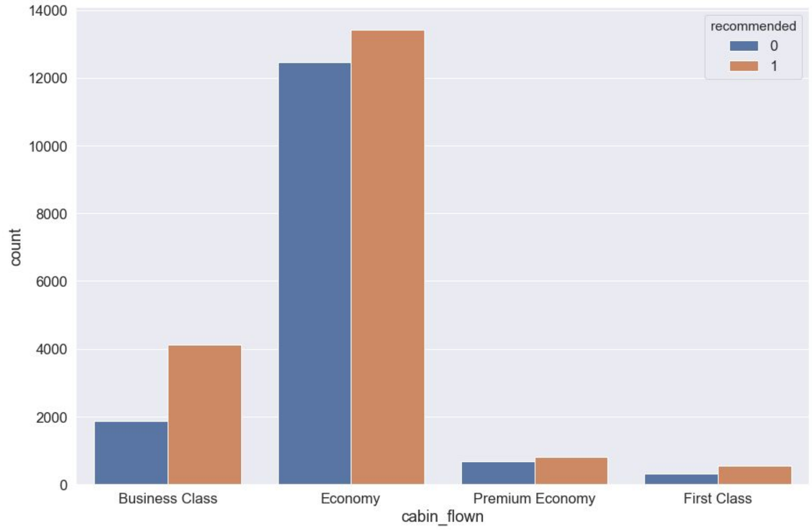
Recommended Rating

From the Recommended rating histogram, we can see that there are more ratings at value 1 than at value 0, which means more travellers recommend an airline they flew with and we can assume that they were overall satisfied with their experience.



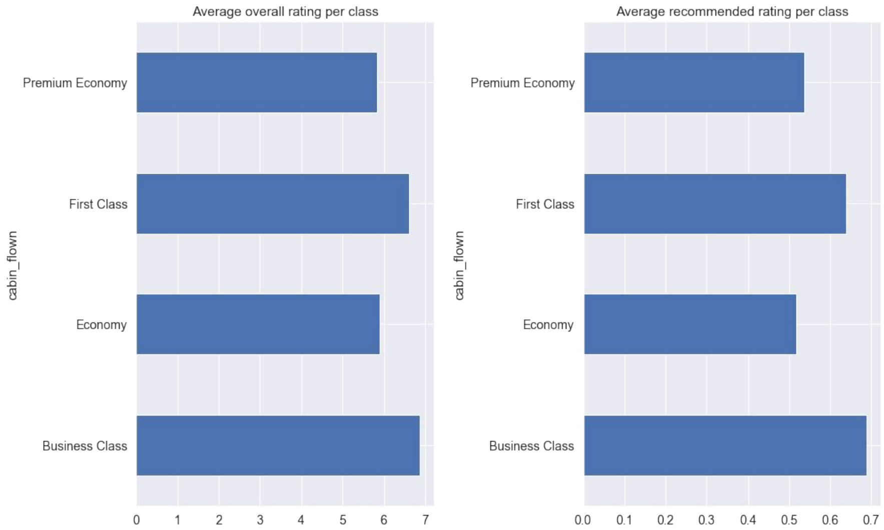
Rating Based on The Flight Class/Cabin Flown

This graph shows us the recommended rating for each flight class. By looking at this graph we can see that there were more travellers that recommended their airline compared to those who didn't recommend it. Especially for Business class there were twice as many for recommendation than not. From this we can assume that travellers in Business class have overall the best experience and are the most satisfied.



Average Overall and Recommend Rating per Class

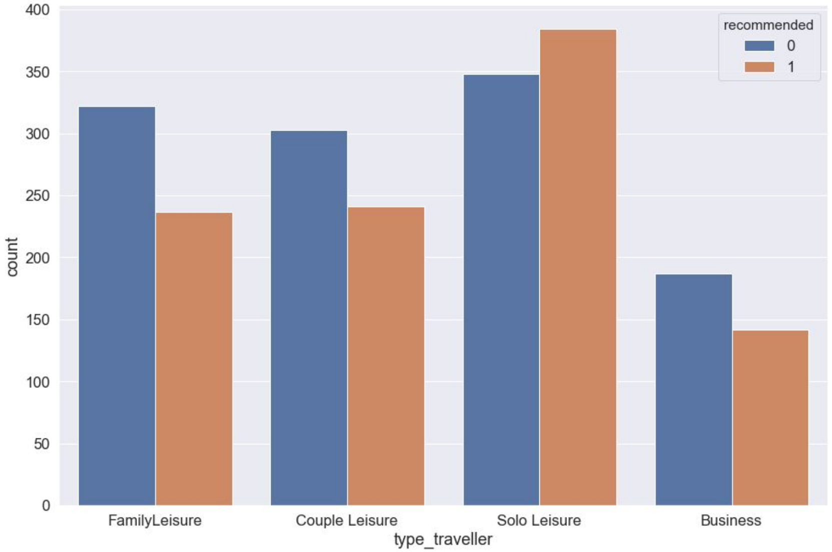
Looking at these two graphs above we can see that Business class has the highest average overall rating as well as average recommended rating, but First class is not far behind. As expected Economy and Premium Economy flight classes have the lowest average rating.



Rating Based on the Traveller Type

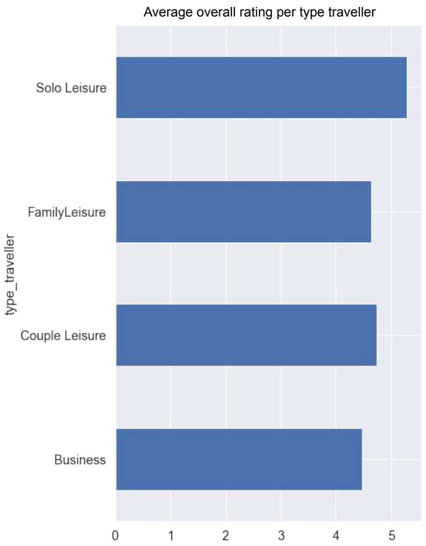
The graph shows recommended ratings for each traveller type. It is a surprise to see that only for travellers that travel individually there are more recommended ratings with yes then no.

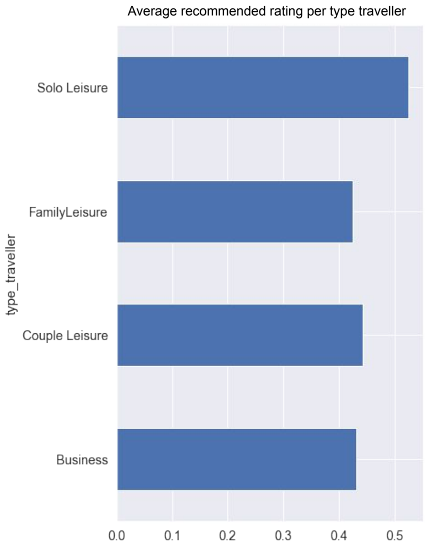
For other types of travellers there are more ratings that would not recommend the airlines they flew with.



Average Overall and Recommend Rating per Type Traveller

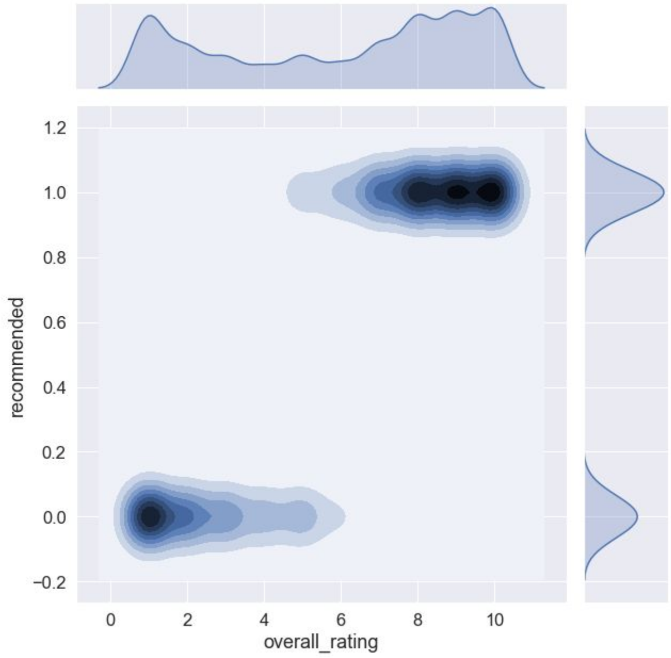
From the above two graphs we can see that Solo type travellers on average gave a higher overall rating and recommended rating than the rest of traveller types. This might be due to not paying attention to details, have no one else with them to worry about. And travellers who travel for a business purpose give on average the lowest ratings.





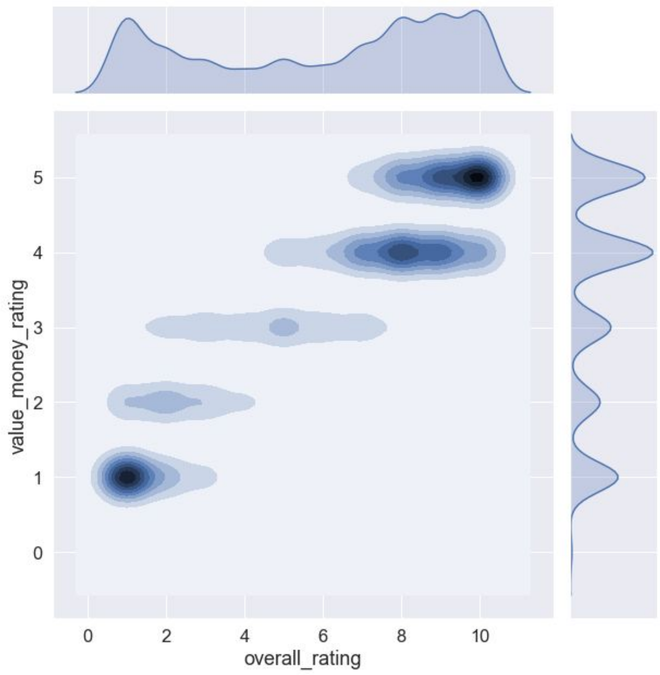
Relationship Between Overall Rating and Recommended Rating

This graph shows us the relationship between overall ratings and recommended ratings. We can see that the limit is at the overall rating value of 6. travellers wouldn't recommend an airline when they gave overall rating closer to value 1, where the highest density of number of not recommended ratings lies. On the other hand, travellers who gave an overall rating of at least 7, would recommend an airline. For recommendation, there are two densities at values 9 and 10 for overall rating.



Rating Based on the Value for Money

We can also observe that there is a high density around value 1 for overall rating and then again between 8 and 10. Similar goes for the "value for money" density graph on the right side. The highest density of ratings are at value 4 and 5 and then again at 1 for low ratings. Since there are not many ratings in the middle we can assume that travellers know if their experience was good or not.



Conclusion

Travellers seem overall either satisfied or very unsatisfied. Which begs the question, does this paint an accurate picture, or are only the travellers on the extremes motivated enough to leave a review? Even if we assume this to be an accurate state of things, there are so many more elements that can influence the rating.

In general, travellers enjoy flying in business class, they give the highest ratings and are most likely to recommend it. It also offers the best value for money. But the vast majority travels in economy class, which is somewhat still more loved than premium economy. The reason for that might be similar to why business class is overall more praised than even first class, the traveller is paying extra for a better experience which probably doesn't live up to their expectations.

Inferential Statistics

One Way ANOVA Results

Both groups have p-value lower than 0.05, which means each one has an independent significant effect on the recommended rating mean value:

	sum_sq	df	F	PR(>F)
C(type_traveller)	6.840535	3.0	9.744814	2.192404e-06
C(cabin_flown)	29.343894	3.0	41.802399	2.995672e-26
Residual	502.140204	2146.0	NaN	NaN

One Way ANOVA Hypothesis for Traveller Type Variable

Null hypothesis *– There is no difference in means of recommended rating between types of travellers.*

Alternative hypothesis *– There is a difference between the means of recommended rating between types of travellers.*

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Business	Couple Leisure	0.0114	0.9	-0.0779	0.1007	False
Business	FamilyLeisure	-0.0076	0.9	-0.0965	0.0812	False
Business	Solo Leisure	0.093	0.0252	0.0081	0.1778	True
Couple Leisure	FamilyLeisure	-0.019	0.9	-0.096	0.0579	False
Couple Leisure	Solo Leisure	0.0816	0.0198	0.0092	0.1539	True
FamilyLeisure	Solo Leisure	0.1006	0.0018	0.0288	0.1724	True

One Way ANOVA Hypothesis for Cabin Flown Variable

Null hypothesis *– There is no difference in means of recommended rating between between flight classes.*

Alternative hypothesis *– There is a difference between the means of recommended rating between flight classes.*

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Business Class	Economy	-0.1705	0.001	-0.1887	-0.1524	True
Business Class	First Class	-0.0505	0.0278	-0.0972	-0.0039	True
Business Class	Premium Economy	-0.1514	0.001	-0.1882	-0.1145	True
Economy	First Class	0.12	0.001	0.0757	0.1644	True
Economy	Premium Economy	0.0192	0.4669	-0.0147	0.0531	False
First Class	Premium Economy	-0.1008	0.001	-0.1556	-0.0461	True

Two Way ANOVA

We also used two-way ANOVA to test for the significance of the interaction between traveller type and cabin flown.

Null hypothesis *– The factors, traveller type and cabin flown, are independent. There is no interaction between them.*

Alternative hypothesis *– The factors, traveller type and cabin flown, are dependent. There is interaction between them.*

Two Way ANOVA Results

The interaction term is not significant ($p > 0.05$). This indicates that there is no interaction effect between the type of traveller and the cabin flown (flight class) on the mean value for the recommended rating. Since this is not significant, the interaction term is to be removed from the model and we can look at the main effects of each variable independently.

	sum_sq	df	F	PR(>F)
C(type_traveller)	6.840535	3.0	9.749635	2.178153e-06
C(cabin_flown)	29.343894	3.0	41.823083	2.930372e-26
C(type_traveller):C(cabin_flown)	2.353190	9.0	1.117980	3.460803e-01
Residual	499.787013	2137.0	NaN	NaN

ANOVA Conclusion

With the ANOVA tests, we discovered that people who travel in Business class are more likely to recommend an airline they flew with. On the other hand, it is interesting to see that people who are travelling for business purpose are less likely to recommend an airline.

Results of the 2-way ANOVA test tells us that there is no interaction between 'type_traveller' and 'cabin_flown' variables, they both have independent influence on the recommended rating.

Machine Learning

Techniques used

*Since this project included text data, as well as ways of determining which class the dependant variable (recommended rating) belongs to, we decided to use the **Multinomial Naive Bayes, Linear Support Vector Classification** and **Random Forest Classifier** to evaluate their 'area under curve' score.*

*To avoid too much of repetition we took advantage of making **Pipelines** and **FeatureUnion**, since our dataset contains different types of features that require different processing pipelines. Our integer columns only needed to be extracted from the dataframe. For string columns we used a **CountVectorizer** class and for text column we used a **TfidfTransformer** class.*

Results for ‘Content’ as Independent Variable

Result from cross-validation of selected classifier

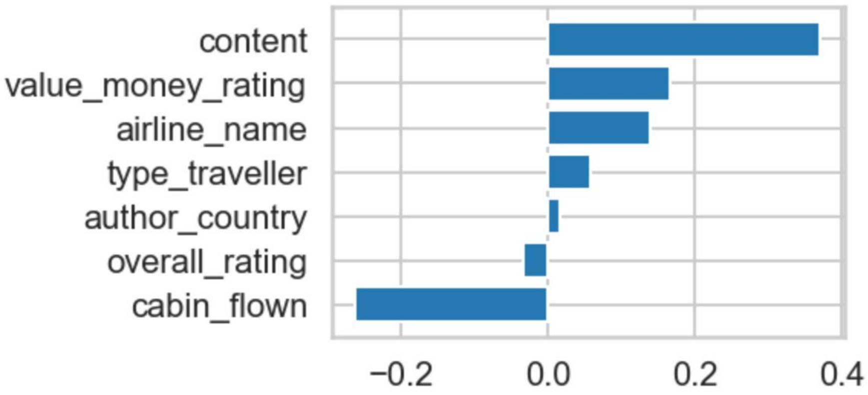
	model_name	fold_idx	roc_auc
0	RandomForestClassifier	0	0.895426
1	RandomForestClassifier	1	0.878161
2	RandomForestClassifier	2	0.891623
3	RandomForestClassifier	3	0.903656
4	RandomForestClassifier	4	0.892179
5	LinearSVC	0	0.948371
6	LinearSVC	1	0.936337
7	LinearSVC	2	0.932137
8	LinearSVC	3	0.933093
9	LinearSVC	4	0.944574
10	MultinomialNB	0	0.928571
11	MultinomialNB	1	0.904952
12	MultinomialNB	2	0.907381
13	MultinomialNB	3	0.937158
14	MultinomialNB	4	0.903234

Results for Good and Bad Words

Because the content of the review is the most important feature we also checked what some of the good and bad words were

```
Good words             P(Yes | word)
             excellent 0.90
           comfortable 0.83
                 great 0.83
              friendly 0.82
                  nice 0.82
                  good 0.81
                  well 0.69
                   new 0.68
                 cabin 0.65
                  crew 0.64
Bad words              P(Yes | word)
                 could 0.41
                   got 0.40
                 delay 0.39
                   get 0.37
                  next 0.36
                   day 0.33
                  hour 0.33
               delayed 0.31
                 never 0.24
                  told 0.15
```

Results for LinearSVC Feature Importance



Once we saw that LinearSVC performed the best, we build few model just with this classifier to see what/which feature are important.

Conclusion

On average the LinearSVC classifier performed better than the RandomForestClassifier or MultinomialNB. But RandomForest was not far behind LinearSVC, especially on smaller sample sizes. Written content is the most important feature, because it gives the most relevant information, compared to ratings which can be very subjective.