Predictions for airline reviews

MAIN QUESTION

Positive reviews are becoming more important in any industry, since an increasing number of customers are basing their purchase decision on them. The airline industry is no different. Airlines are putting in a lot of effort to offer good value without overextending on the business cost. Understanding customer feedback is an important component of that, but offers only a snapshot of the current situation. Would it be possible for them to predict future customer reviews based on the current situation, and better anticipate potential issues?

PROBLEM

Airlines are continuously tuning and balancing passenger experience versus operational costs. They 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? How can they catch service related dissatisfactions before they turn into a negative news story?

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.

Long lines due to security procedures at check-in, cramped seating, inconvenient schedules, poor service - the list of airline travelers' complaints is a lengthy one. The perception that air travel is an ordeal makes it very difficult for airlines to charge the higher prices that are necessary to return to profitability. Social media has propelled a number of what can only be described as PR disasters recently, and undoubtedly caused harm to the industry.
- From Investopedia

CLIENTS

AIRLINES
Airline can track customer satisfaction with the offered services and predict how it might change going forward, thus being able to catch and resolve issues early on. It would also enable them to focus on the parts of the service passenger care about the most. As well as explore hypothetical situations where one part of the service would be unsatisfactory and observe the effect that would have on potential future reviews.

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.

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.

DATA

To better understand the ratings and reviews passengers have left the airlines we need the following data:

- Airline name
- Country of passenger
- Date of review
- Written review content
- Cabin class
- Ratings
    - Overall rating
    - Food rating
    - Seat comfort
    - Cabin staff
    - In-flight entertainment
    - Value for money
    - Wi-Fi connectivity
    - Ground service
- Would the passenger recommend this Airline to others

We will be basing our models on a Skytrax Reviews dataset we discovered on Github. It already contains everything we need - https://github.com/rbackupX/skytrax-reviews-dataset

APPROACH

We would start by getting an understanding of the data. There are two methods we could use for the prediction part, depending on what it is we want to predict. We would use the classification method for predicting if a passenger would recommend that airline to others or not. For predicting the overall rating, which is on a scale from 1 to 10, we would use regressions.

DELIVERABLES

- Code
- Report
- Visualizations
- Github repository with all the documentation

DATA WRANGLING

There was not much of data cleaning and transformation that had to be done. The data set was already in a tidy format. The only thing that we changed was ‘FamilyLeisure’ in the ‘TypeTraveller’ column, by adding spaces between words to make ‘Family Leisure’ instead, and make it consistent across categories.

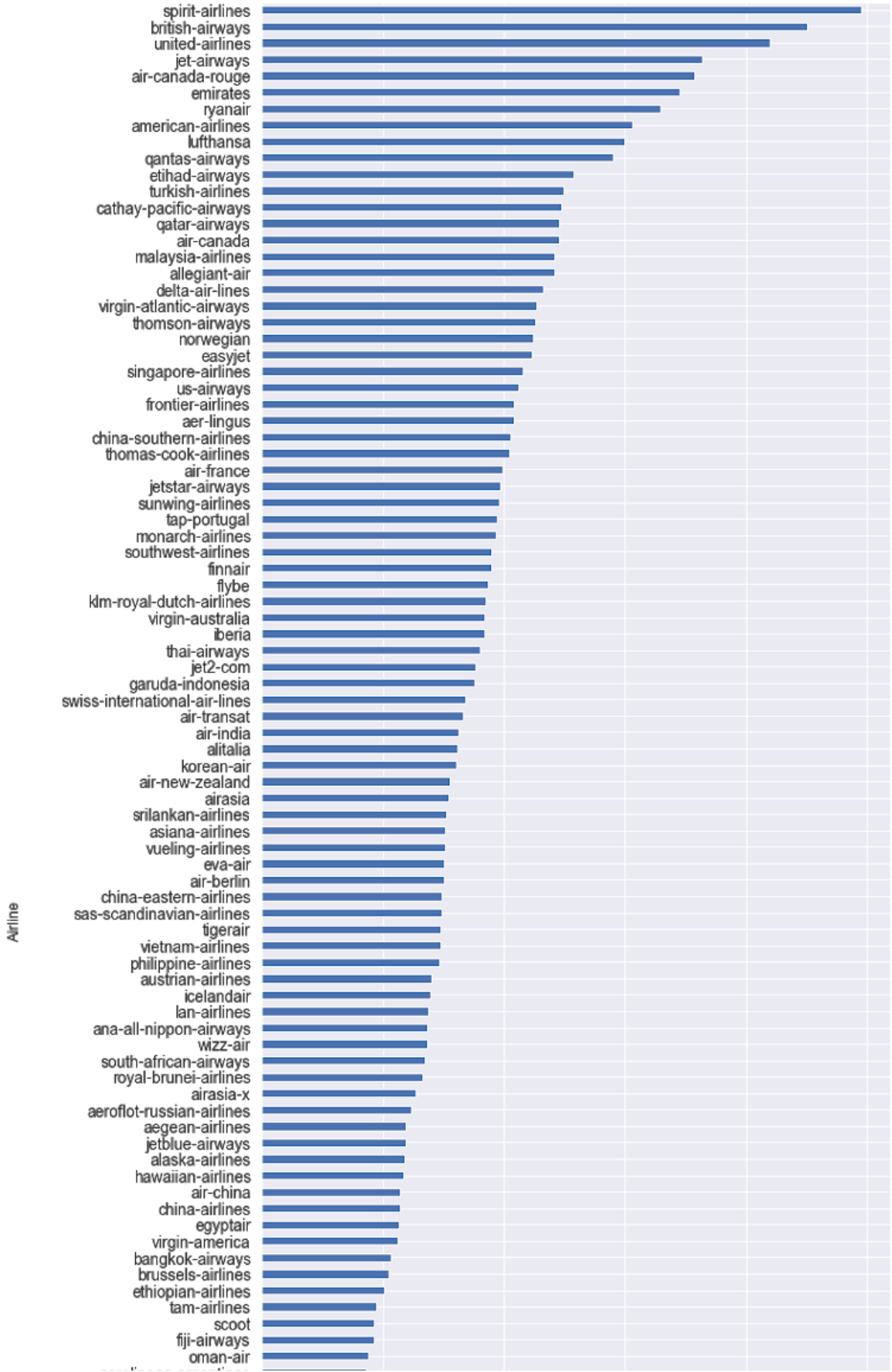
DATA STORY

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. Between these airlines there are large differences in the number of reviews, for example Spirit airlines has 990 reviews, which is an order of magnitude more than the lowest number we have chosen. After filtering out the airlines with fewer than 100 review, we are left with 35609 rows and 20 columns, which is around 29,1 MB.

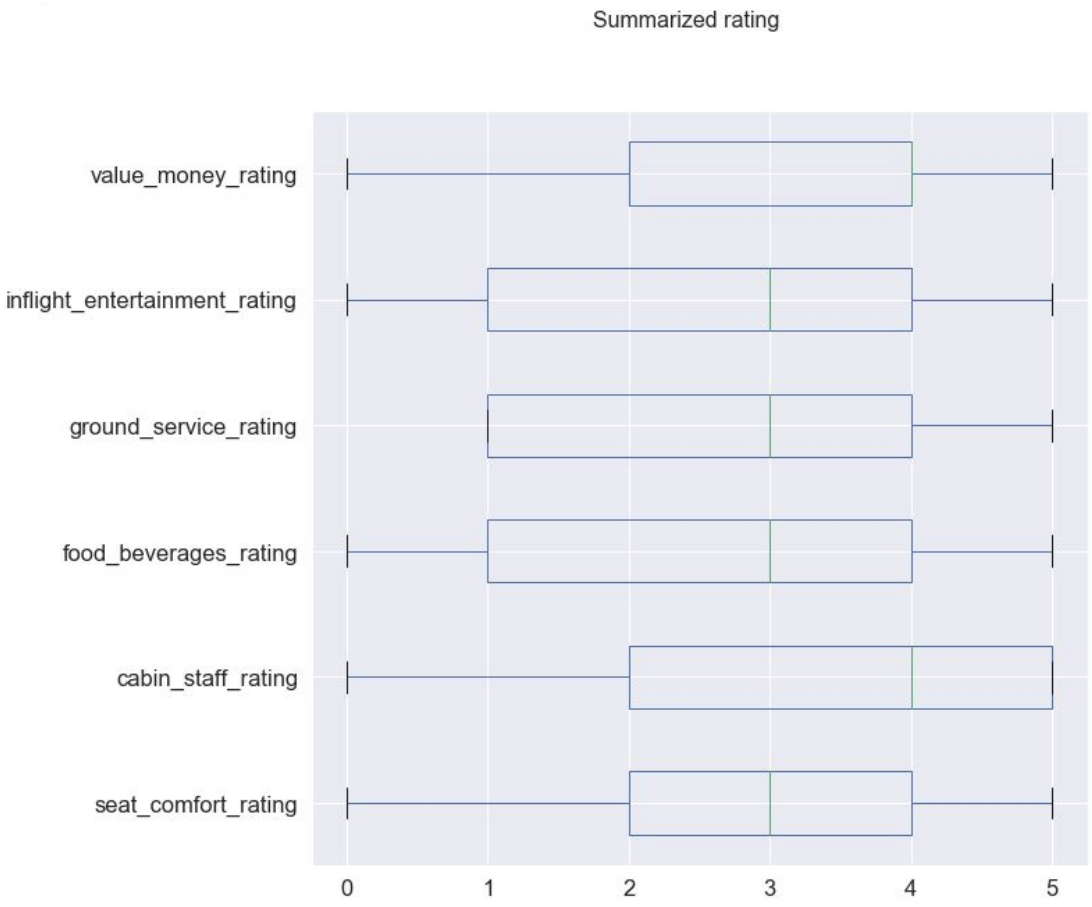
Here are some of the questions we wanted to gain a deeper understanding of:

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? Who will be more likely to give a better rating?
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?

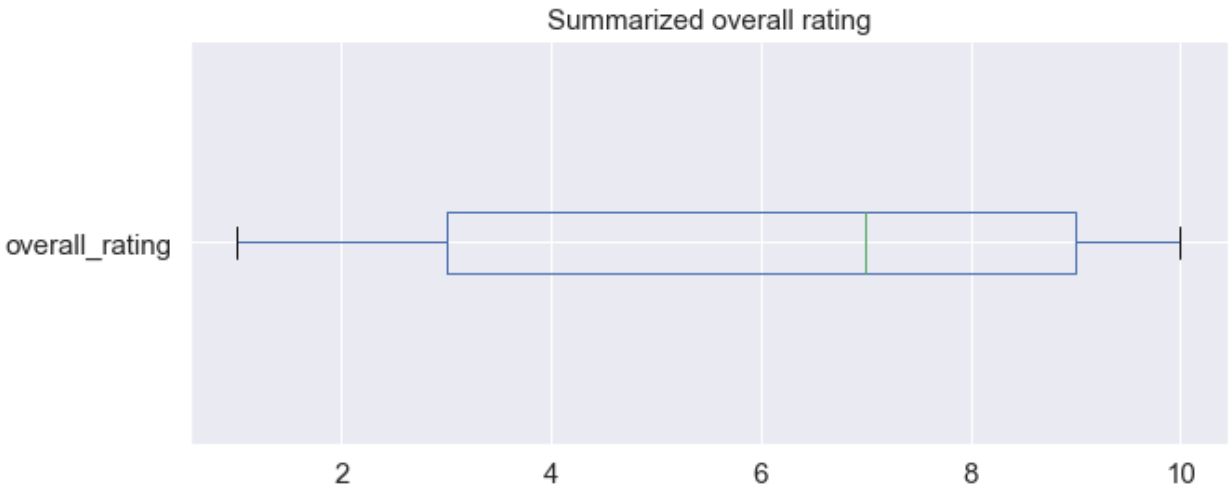
DATA SUMMARY

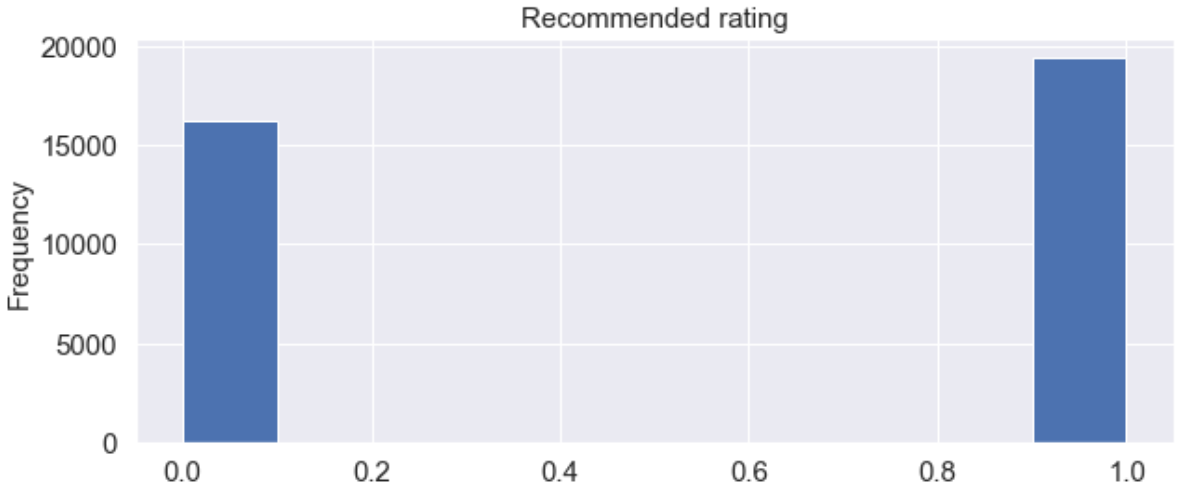

Airline
spirit-airlines
british-airways
united-airlines
jet-airways
air-canada-rouge
emirates
ryanair
american-airlines
lufthansa
qantas-airways
etihad-airways
turkish-airlines
cathay-pacific-airways
qatar-airways
air-canada
malaysia-airlines
allegiant-air
delta-air-lines
virgin-atlantic-airways
thomson-airways
norwegian
easyjet
singapore-airlines
us-airways
frontier-airlines
aer-lingus
china-southern-airlines
thomas-cook-airlines
air-france
jetstar-airways
sunwing-airlines
tap-portugal
monarch-airlines
southwest-airlines
finnair
flybe
klm-royal-dutch-airlines
virgin-australia
iberia
thai-airways
jet2-com
garuda-indonesia
swiss-international-air-lines
air-transat
air-india
alitalia
korean-air
air-new-zealand
airasia
srilankan-airlines
asiana-airlines
vueling-airlines
eva-air
air-berlin
china-eastern-airlines
sas-scandinavian-airlines
tigerair
vietnam-airlines
philippine-airlines
austrian-airlines
icelandair
lan-airlines
ana-all-nippon-airways
wizz-air
south-african-airways
royal-brunei-airlines
airasia-x
aeroflot-russian-airlines
aegean-airlines
jetblue-airways
alaska-airlines
hawaiian-airlines
air-china
china-airlines
egyptair
virgin-america
bangkok-airways
brussels-airlines
ethiopian-airlines
tam-airlines
scoot
fiji-airways
oman-air

The majority of airlines on our list have between 200 and 500 reviews. From all the airlines on our list of 112, Spirit dominantes in the number of reviews. The popularity of the airline might be due to the ultra low cost, but that also brings them to the bottom of our list when it comes overall rating.



All features in the graph above have a rating from 0 to 5, where 0 is the worst and 5 the best. We can see that overall Cabin staff has the best rating. Inflight entertainment and Food and beverage have the worst ratings, they both have minimum rating 0 and the middle 50% of the rating is in the range between 1 and 4. Ground service has a similar range for 50% of ratings in the middle, but what is unexpected is that its minimum value is 1.

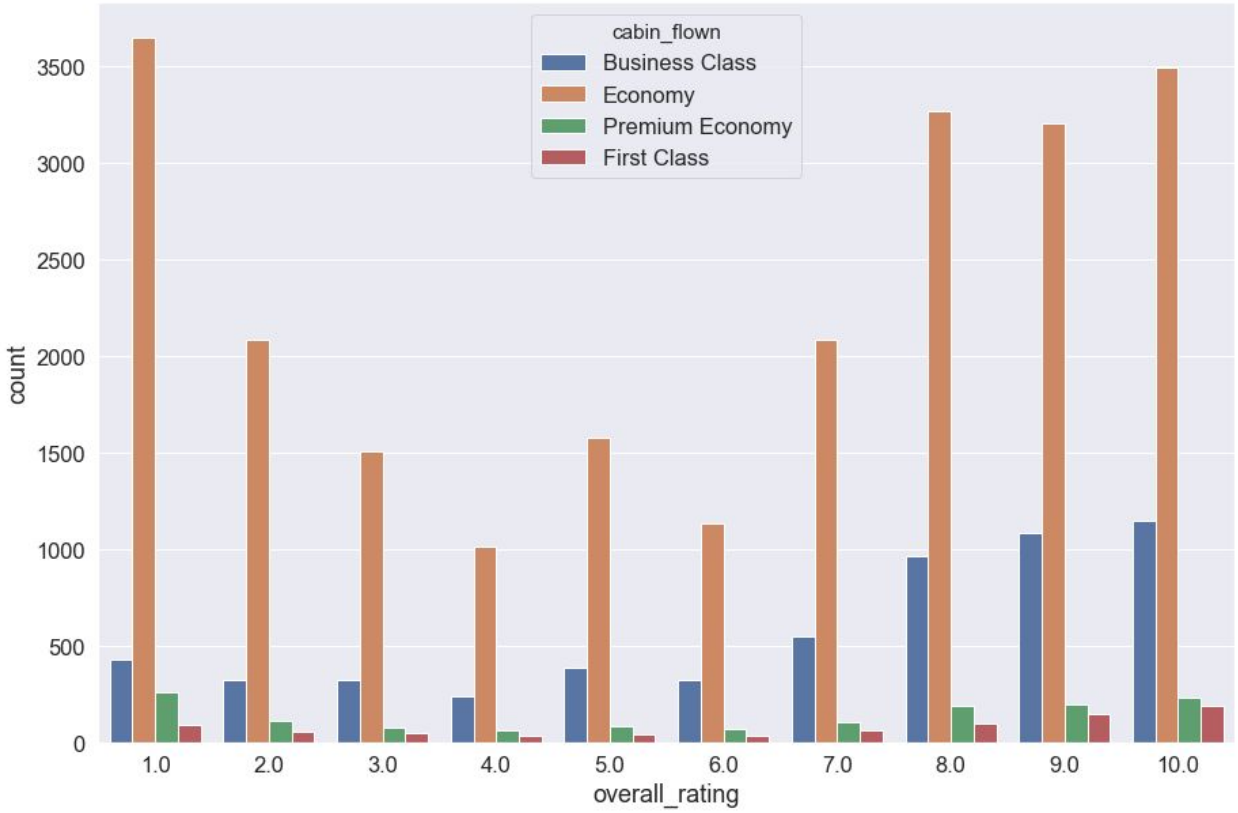




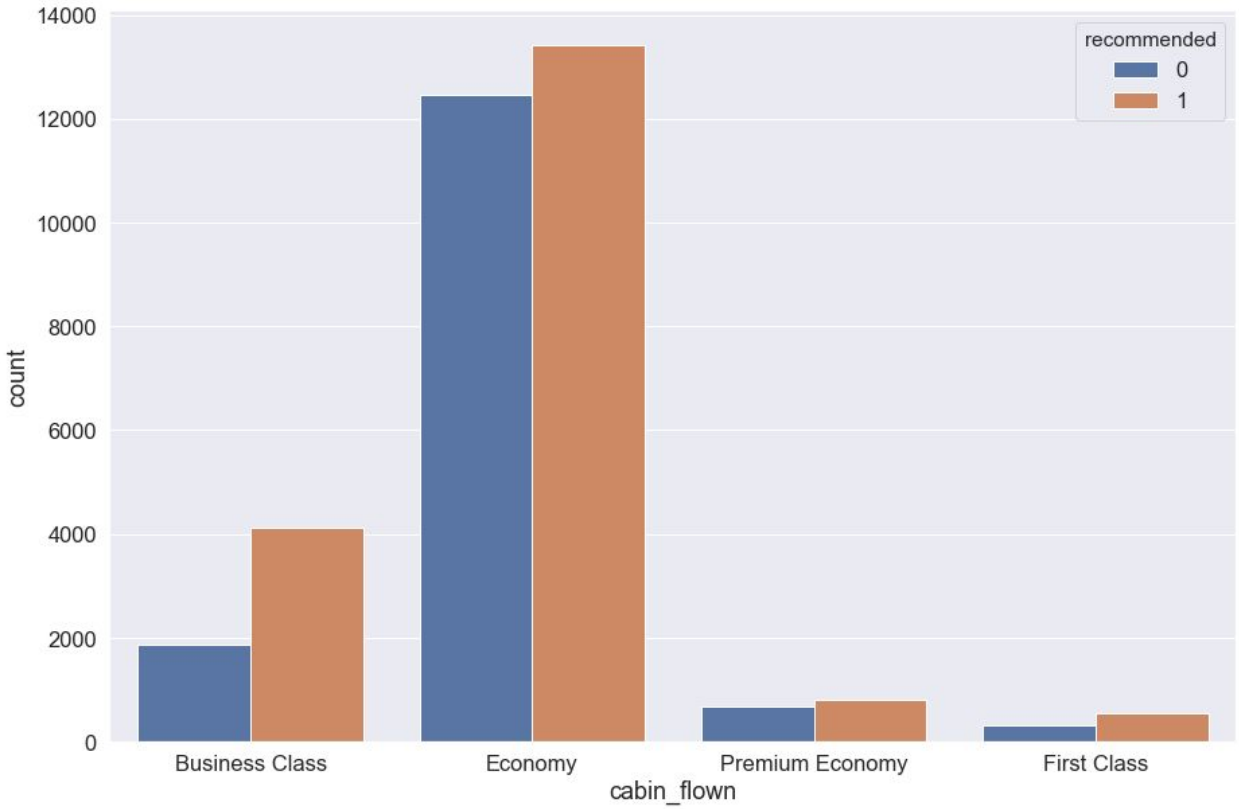
Overall rating has values between 1 and 10, where 1 is the worst and 10 the best. These two values also represent minimum and maximum ratings. And the 50% of ratings is between 3 and 9, with the mediana at 7. From this we can assume that travellers have an overall good experience. Let's see if that means that they would recommend it to others as well.

From the Recommended rating histogram, we can see 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.

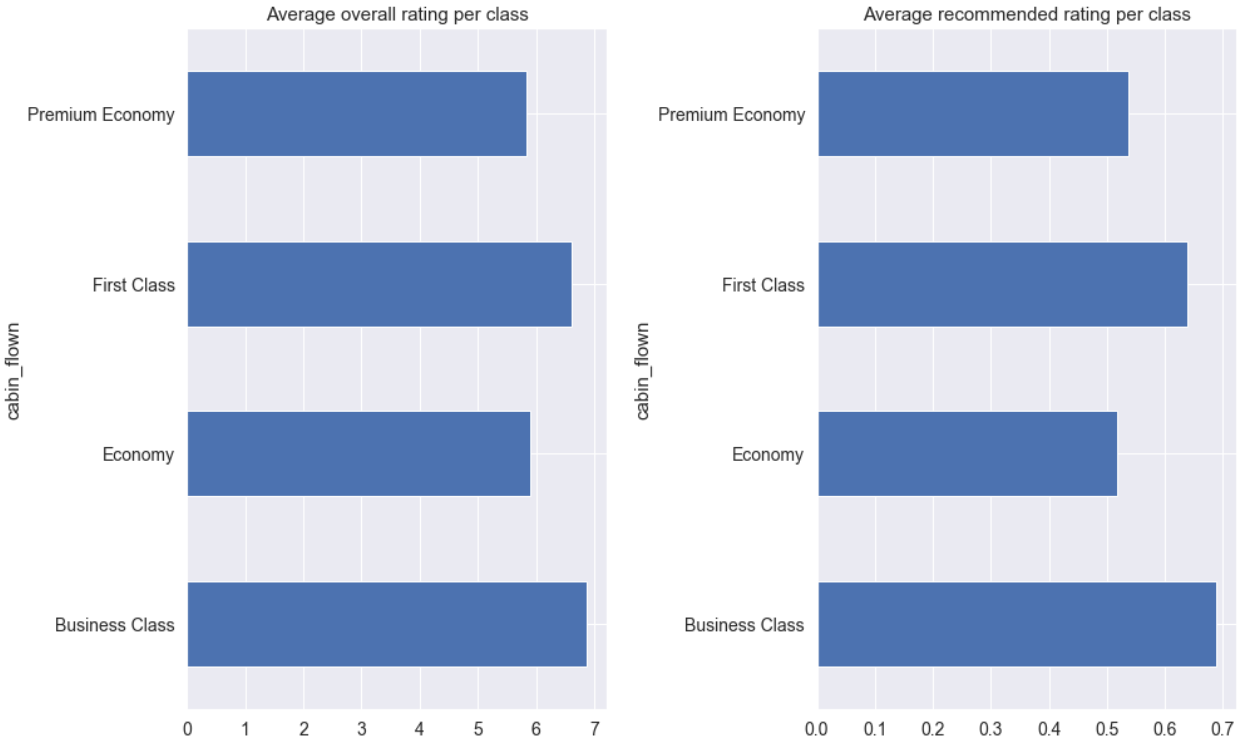
RATING BASED ON THE FLIGHT CLASS/CABIN FLOWN



The above graph shows the distribution of the data for Overall rating. What we can see there are the most ratings for Economy class and the least for First class. Which is no surprise, since the Economy class has more seats available per flight than First class, if there even is a First class available. We can also notice a similar pattern occurring across all the classes, with a lot of ratings between 8 and 10, a dip in the middle and a peak at 1. Premium economy ratings follow closer to Economy class, while First class are similar to Business class ratings. Which all makes sense.

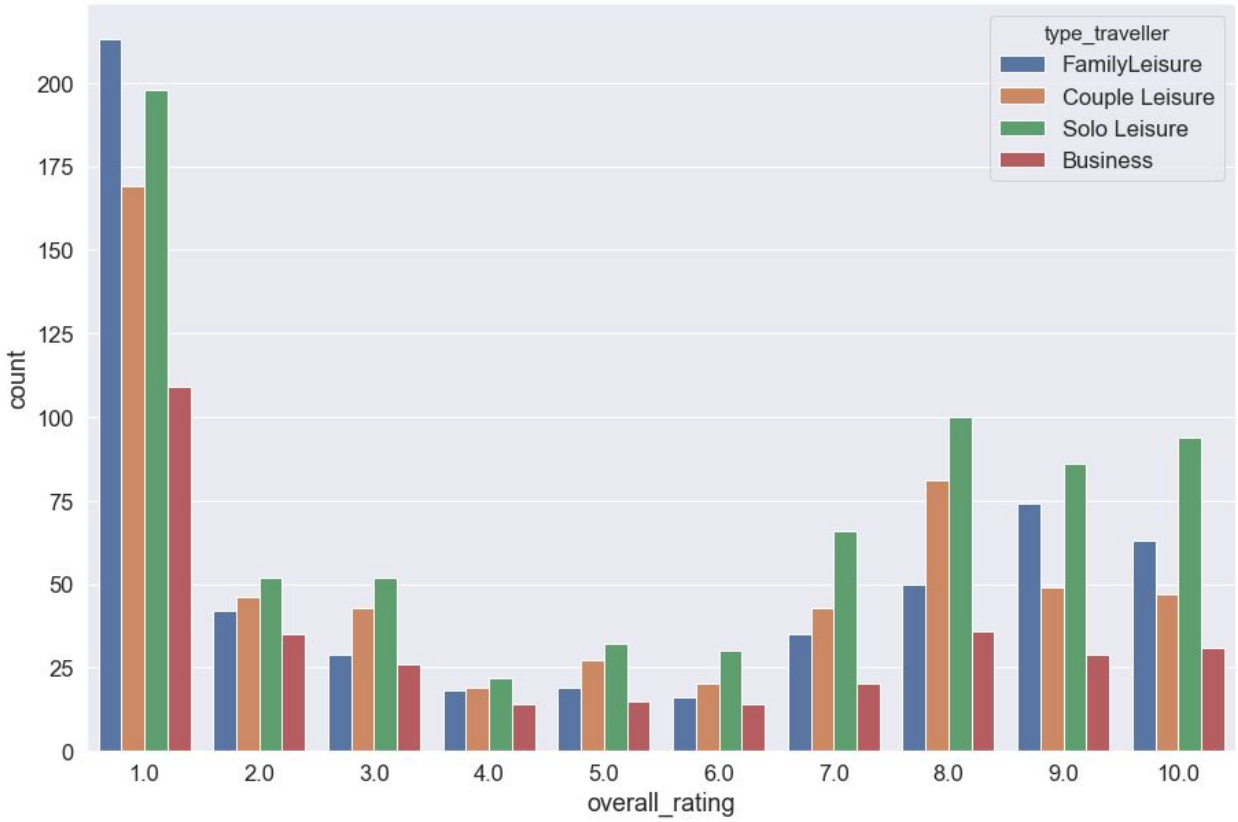


This graph shows us the recommended rating for each flight class. By looking at this graph we can see that there were more travellers that recommended their airline compared to those who didn't recommend it. Especially for Business class there were twice as many for recommendation than not. From this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

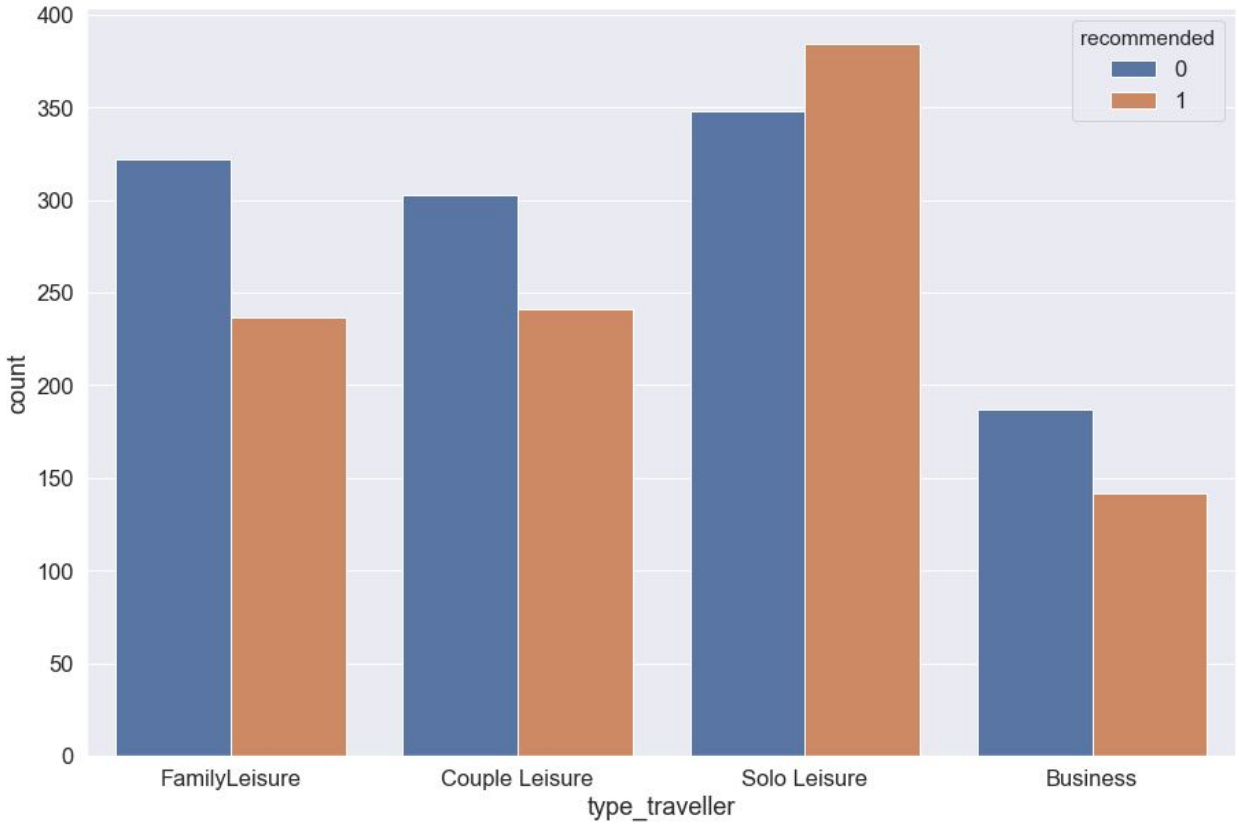


Looking at these two graphs above we can see that Business class has the highest average overall rating as well as average recommended rating, but First class is not far behind. As expected Economy and Premium Economy flight classes have the lowest average rating.

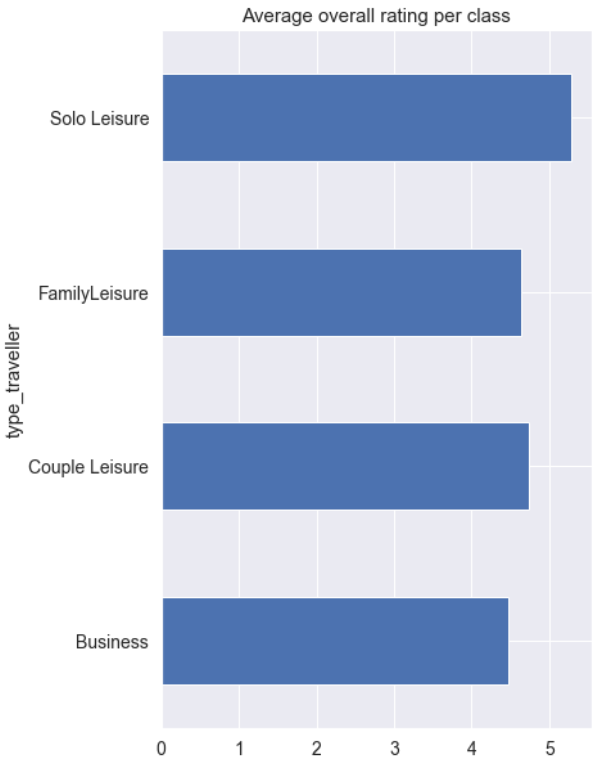
RATING BASED ON THE TRAVELLER TYPE

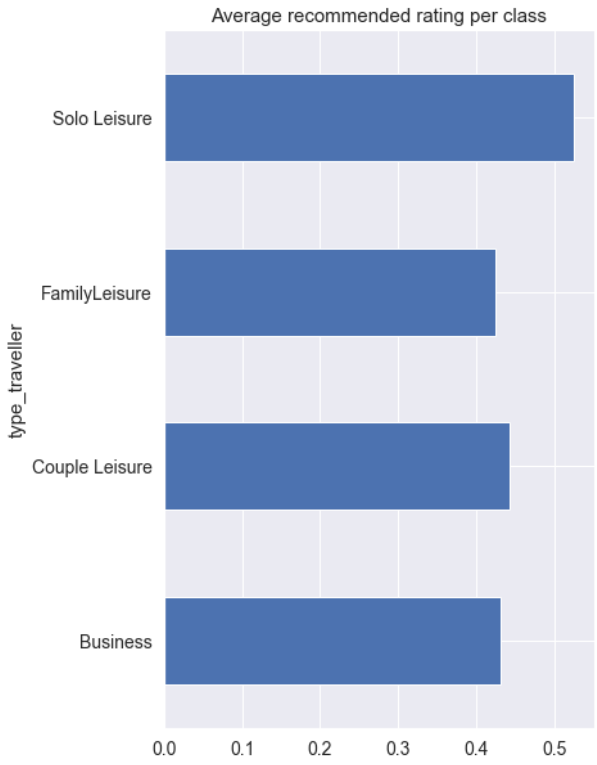


The above graph shows us the distribution of the overall rating for each traveller type. We can see that each traveller type has the most rating for value 1, which means they weren't satisfied with their flight. It is interesting to see that the lowest number of ratings have values 4, 5 and 6. From this we can assume that travellers are not indecisive, and have a strong sense if they had good or bad experience.



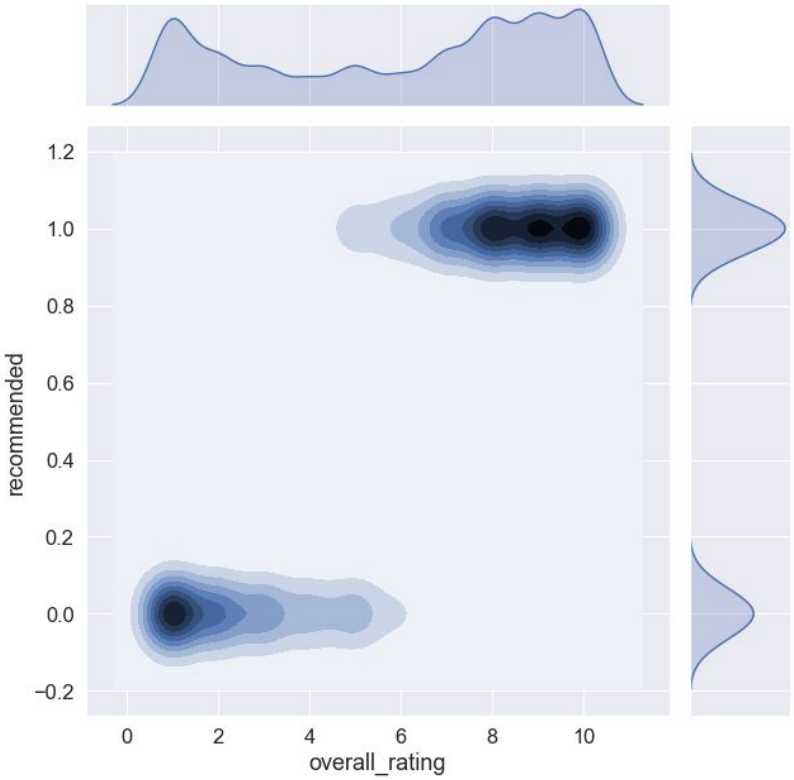
The above graph shows recommended ratings for each traveller type. It is a surprise to see that only for travellers that travel individually there are more recommended ratings with yes then no. For other types of travellers there are more ratings that would not recommend the airlines they flew with.





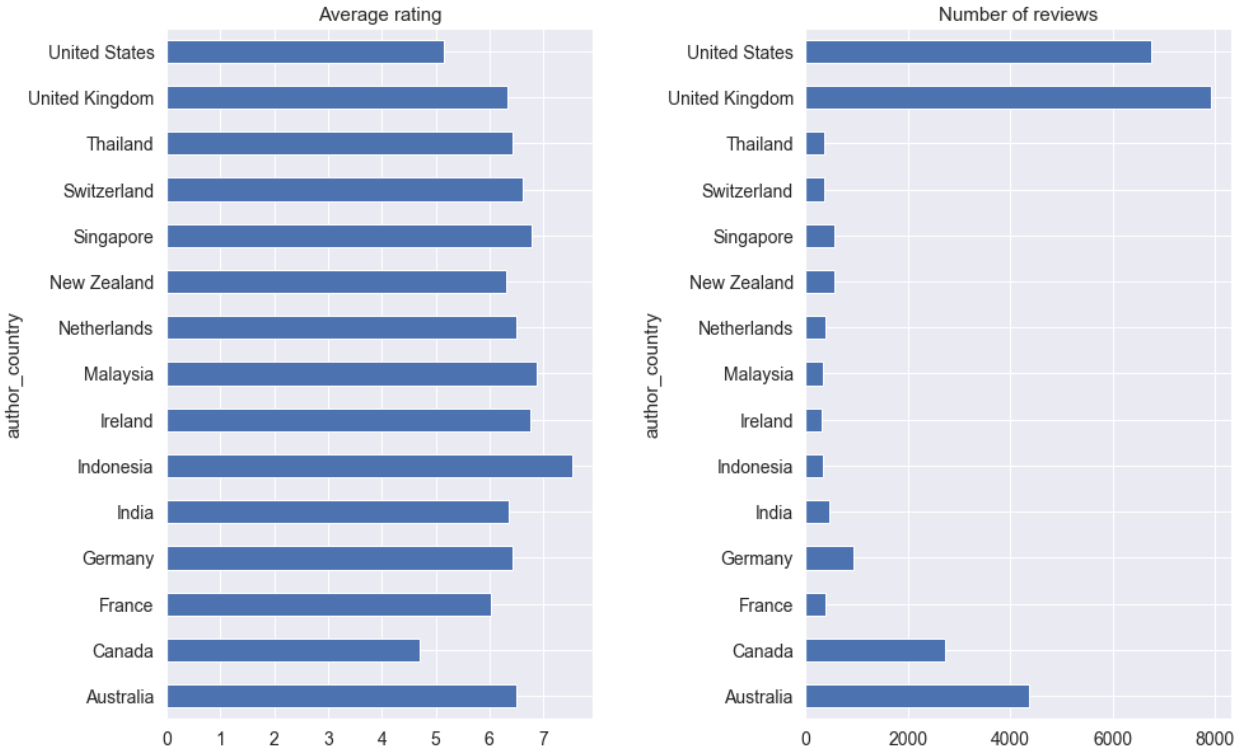
From the above two graphs we can see that Solo type travellers on average gave a higher overall rating and recommended rating than the rest of traveller types. This might be due to not paying attention to details, have no one else with them to worry about. And travellers who travel for a business purpose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

RELATIONSHIP BETWEEN OVERALL RATING AND RECOMMENDED RATING

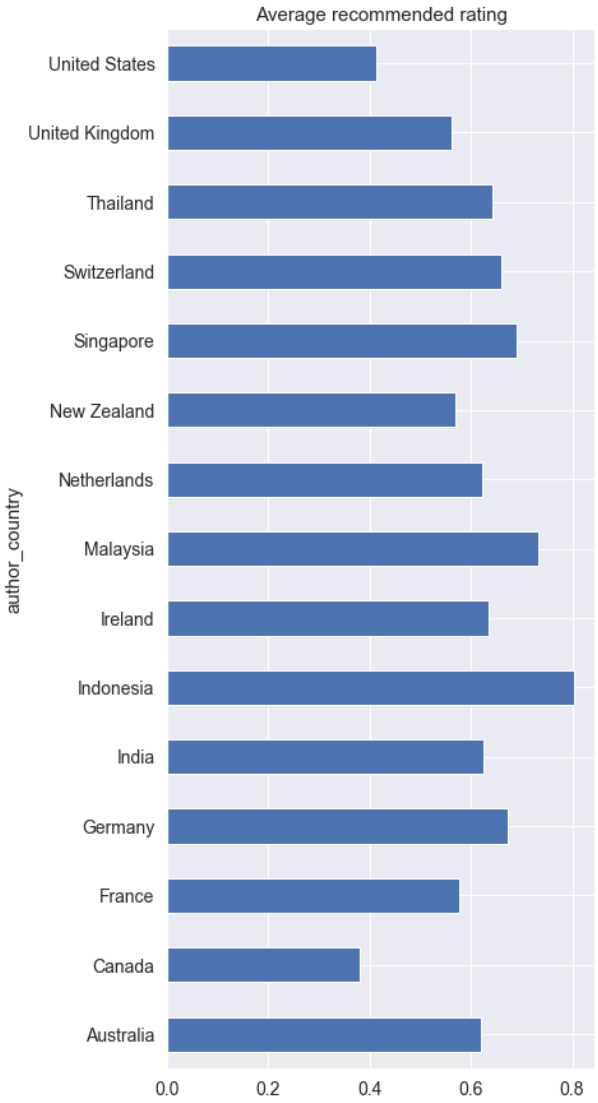


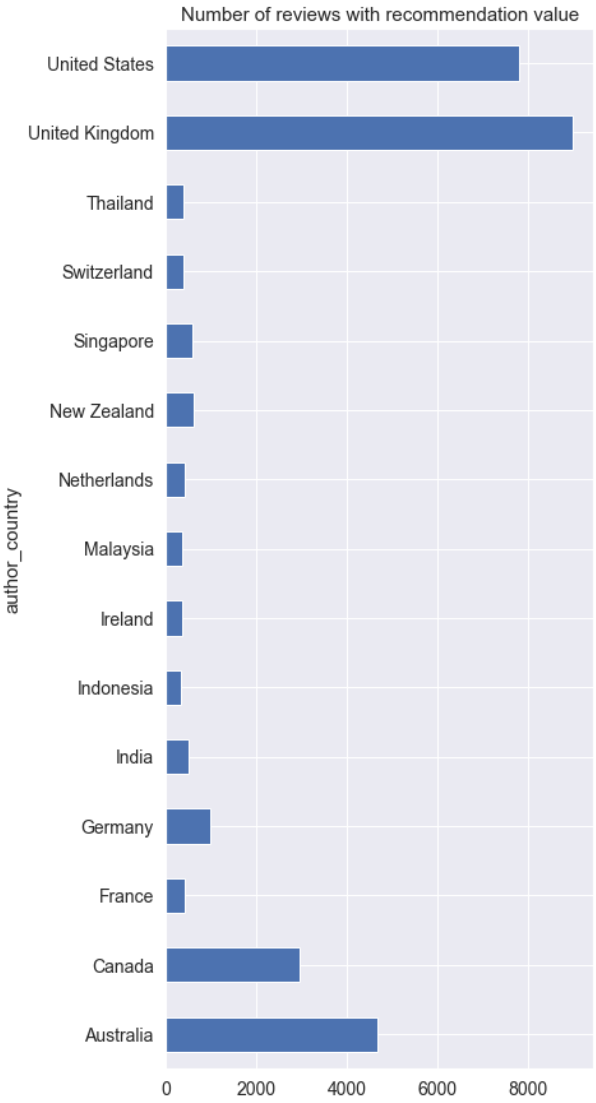
This graph shows us the relationship between overall ratings and recommended ratings. We can see that the limit is at the overall rating value of 6. travellers wouldn't recommend an airline when they gave overall rating closer to value 1, where the highest density of number of not recommended ratings 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 TRAVELLER'S COUNTRY



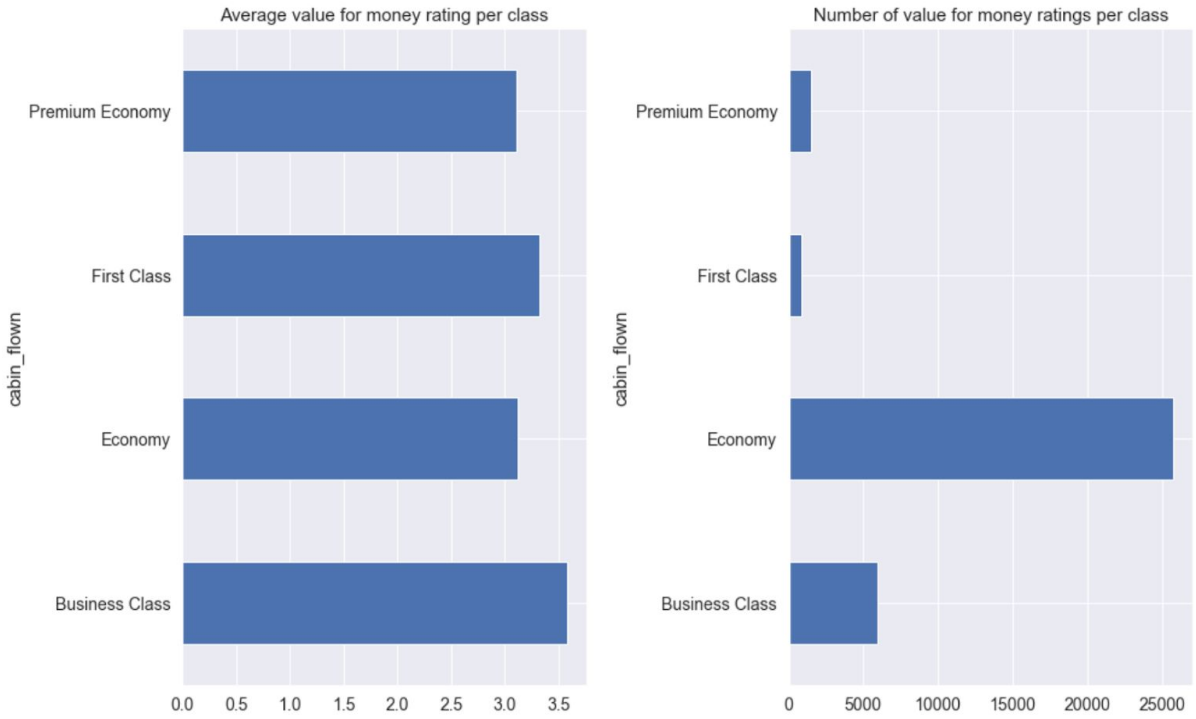
By looking at the right graph first, we could group countries in 3 groups by number of reviews for easier comparison. The USA and UK in one, AUT and CAN in second and the rest in third. We can see that countries with number of reviews lower than 2000 have high average Overall rating. On the other hand the UK is not far behind, even though it has the highest number of reviews. It is unexpected to see that travellers from Canada aren't satisfied with their flight experience compared to the others.



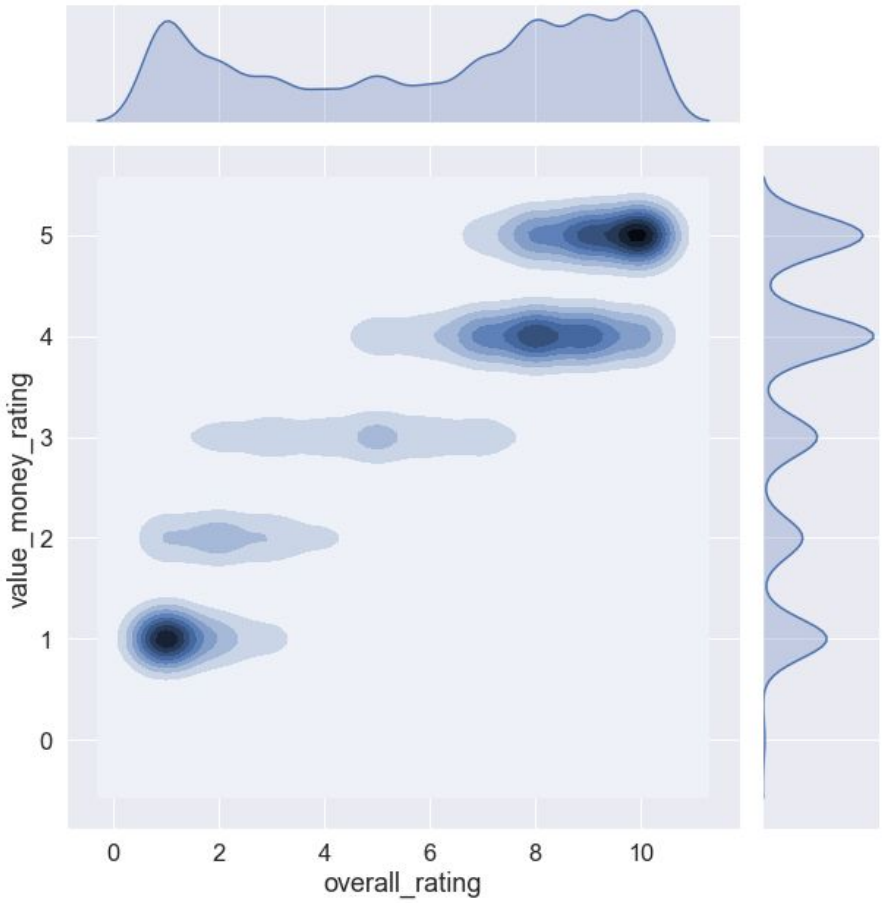


In this two graphs we can see a similar pattern as above. travellers from Canada are on average less likely to recommend their airline then the rest. travellers from the countries with less than 2000 reviews are more likely to be recommending. And the UK is not that far away, taking into consideration that it has the highest number of reviews.

RATING BASED ON THE VALUE FOR MONEY



By looking at the above two graphs seems that travellers flying in Business class get better value for money. As expected, in Economy class travellers on average gave the lowest value for money rate.



From this graph we can see the relationship between the overall rating and value for money rating. For each “value for money” rating we can see one density area that belongs to the overall rating. Traveller who give 1 for “money for value”, often gave overall rating value of 1. When they gave 10 for the overall rating, then the “value for money” rating is at 5.

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 OF DATA STORY

Overall what we have seen from the graphs and data is that travellers are 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 that probably doesn't live up to their expectations.

Things that also have a measurable effect is the type of travel. Usually people who travel solo for leisure are way happier with their airlines. Compared to people travelling with families, who would not recommend an airline, even if they give it a good overall rating. Looks like travelling with others makes people a bit more miserable, even couples aren't as happy as solo travellers. Flying for work also isn't the most favourite thing for many, but flying in business class compared to economy makes the trip a bit more tolerable.

And then there is the country of origin, which leaves a lot of questions. At a glance it looks like people from less developed countries leave better review ratings. Is that due to the difference in living standards, or are cultural differences at play?

Most of the data seems mostly to make sense. The thing that sticks out is that the cabin staff receives higher ratings than food, entertainment, and seat comfort. This should give airlines a good sense of where to invest a bit more.

INFERENTIAL STATISTICS

ONE-WAY ANOVA

For our statistical analysis we decided to use the one-way ANOVA (analysis of variance), to compare the means of groups by analyzing variances. We have 4 types of travellers (Solo, Couple, Family and Business), and 4 flight classes (Economy, Premium Economy, Business and First), which represent groups in our ANOVA analysis. Traveller type and cabin flown are independent variables. Data for recommended rating represents a dependable variable.

Results from ANOVA table:

	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

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

Post-hoc Testing

We used Tukey HSD post-hoc comparison test between different group means. And the results are in the following tables.

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

Group1 and group2 columns are the groups being compared, the meandiff column is the difference in means of the two groups being calculated as group2 – group1, the lower/upper columns are the lower/upper boundaries of the 95% confidence interval, and the reject column states whether or not the null hypothesis should be rejected.

We can reject the null hypothesis based on three comparisons: There is a difference in means for recommended rating between Business and Couple Leisure traveller type. There is also a difference in means for recommended rating between Business and Family Leisure traveller type. We can also reject for the comparison between Couple Leisure and Family Leisure traveller type.

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

From the comparison test, we can reject the null hypothesis for the comparison between Economy and Premium Economy class. Which means there is not a statistically significant difference in the mean values for recommended rating between these two classes.

Other tests

We used Levene's test to check for homogeneity of variance.

Levene's test for homogeneity of variance is not statistically significant for the traveller type variable (statistic=1.3126570838357872, pvalue=0.26853625706562273), which indicates that the traveller type groups have equal variances.

Levene's test for homogeneity of variance is statistically significant for cabin flown (statistic=202.47664595260858, pvalue=3.5377794911393746e-130), which indicates that the flight classes don't have equal variances.

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.

Results from ANOVA table:

	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

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.

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.

For each classifier, we used the same data sample, which we split into a training and test set (70/30). We experimented with sample sizes and features selection, based on the available non null values of features. 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.

We used AUC scores on our testing dataset as performance measure of our model. This measured score is used for binary classification task as we only care for the final class prediction and don’t want to tune the threshold. This project is far more detailed than it is outlined in this report. For a deeper analysis with all the models and graphs, check out the file on Github.

RESULTS - FOR EACH TECHNIQUE

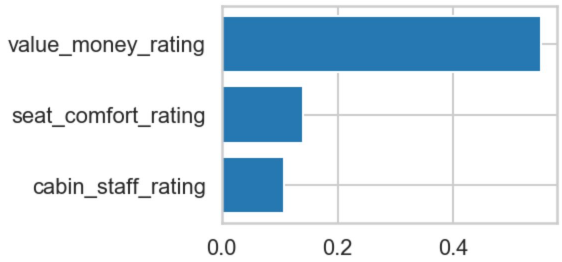
On average Linear Support Vector Classification performed the best. For our dependable variable we selected the ‘recommended’ column, and experimented with various combinations of independent variables to see how they affect the recommendation rating. This project is far more detailed than it is outlined in this report. All results are in the table below.

CLASSIFIER	DATA SIZE (# rows)	FEATURES	AUC SCORE	CONFUSION MATRIX
MultinomialNB	35609	Text content	0.8474	true positive: 4269, true negative: 3305, false positive: 800, false negative: 529
MultinomialNB (with best parameters)	35609	Text content	0.7772	true positive: 11226, true negative: 8259, false positive: 3113, false negative: 2328
LinearSVC	100	Text content	0.65687	true positive: 55, true negative: 15, false positive: 28, false negative: 2
RandomForestClassifier	100	Text content	0.6044	true positive: 53, true negative: 12, false positive: 31, false negative: 4
LinearSVC	35609	Text content	0.8846	true positive: 5199, true negative: 4258, false positive: 594, false negative: 632
RandomForestClassifier	35609	Text content	0.8153	true positive: 4748, true negative: 3961, false positive: 891, false negative: 1083
LinearSVC	100	value_money_rating, seat_comfort_rating, cabin_staff_rating	0.95	true positive: 20, true negative: 9, false positive: 1, false negative: 0
RandomForestClassifier	100	value_money_rating, seat_comfort_rating, cabin_staff_rating	0.950	true positive: 20, true negative: 9, false positive: 1, false negative: 0
LinearSVC	100	airline_name, cabin_flown	0.5	true positive: 20, true negative: 0, false positive: 10, false negative: 0
RandomForestClassifier	100	airline_name, cabin_flown	0.5	true positive: 20, true negative: 0, false positive: 10, false negative: 0
LinearSVC	100	airline_name, cabin_flown, author_country	0.55	true positive: 20, true negative: 1, false positive: 9, false negative: 0

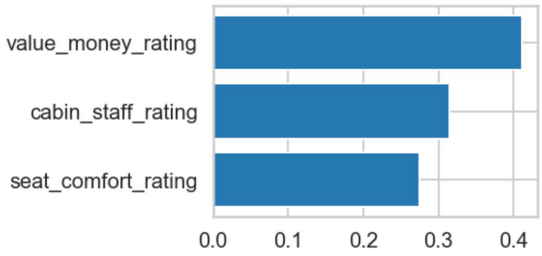
RandomForestClassifier	100	airline_name, cabin_flown, author_country	0.525	true positive: 19, true negative: 1, false positive: 9, false negative: 1
LinearSVC	100	seat_comfort_rating, cabin_staff_rating, value_money_rating	0.9	true positive: 18, true negative: 9, false positive: 1, false negative: 2
RandomForestClassifier	100	seat_comfort_rating, cabin_staff_rating, value_money_rating	0.9	true positive: 18, true negative: 9, false positive: 1, false negative: 2
LinearSVC	1148	airline_name, author_country, aircraft, type_traveller, cabin_flown	0.6068	true positive: 181, true negative: 49, false positive: 74, false negative: 41
RandomForestClassifier	1148	airline_name, author_country, aircraft, type_traveller, cabin_flown	0.6132	true positive: 182, true negative: 50, false positive: 73, false negative: 40
LinearSVC	34138	airline_name, cabin_flown	0.6353	true positive: 4381, true negative: 2274, false positive: 2316, false negative: 1271
RandomForestClassifier	34138	airline_name, cabin_flown	0.6326	true positive: 4353, true negative: 2272, false positive: 2318, false negative: 1299
LinearSVC	34138	airline_name, cabin_flown, content	0.8906	true positive: 5105, true negative: 4030, false positive: 560, false negative: 547
RandomForestClassifier	34138	airline_name, cabin_flown, content	0.8245	true positive: 4736, true negative: 3723, false positive: 867, false negative: 916
LinearSVC	30859	author_country, cabin_flown, overall_rating, author_country, value_money_rating, content	0.9413	true positive: 5385, true negative: 3361, false positive: 270, false negative: 242
RandomForestClassifier	30859	author_country, cabin_flown, overall_rating, author_country, value_money_rating, content	0.9201	true positive: 5208, true negative: 3321, false positive: 310, false negative: 419

LinearSVC	2150	airline_name, author_country, content, type_traveller, cabin_flown, overall_rating, value_money_rating	0.9446	true positive: 277, true negative: 332, false positive: 21, false negative: 15
RandomForestClassifier	2150	airline_name, author_country, content, type_traveller, cabin_flown, overall_rating, value_money_rating	0.9141	true positive: 265, true negative: 325, false positive: 28, false negative: 27

We can notice that the AUC score is not only dependent on the size of the data, but also on the selected independent features. If we compare the LinearSVC model with only one independent feature (text content) and the LinearSVC model with 7 independent features, we can see that the latter performed much better. It was unexpected to see that we got the best results from both classifiers with 3 independent features (ratings), both with an AUC score of 0.95, even though their sample size was small, only 100 rows. For these two classifiers we also checked feature importance:

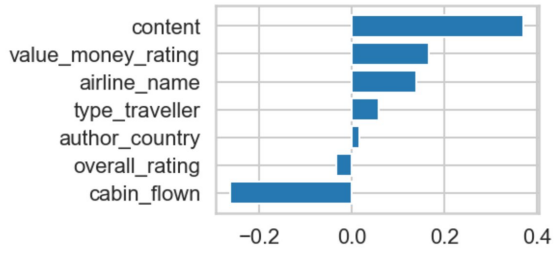


LinearSVC

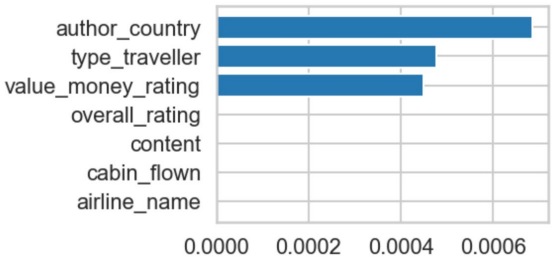


RandomForestClassifier

Let's take a look at the feature importance for the models with the most features included.



LinearSVC



RandomForestClassifier

Since the LinearSVC classifier performed better, looking at this feature importance we can say that what travellers write in reviews significantly affects the recommendation. The second importance feature is ‘value for money’, we can assume that people care about their experience and not just the price of the ticket. It is interesting to see that the recommendation depends more on traveller type - travelling alone, as a couple, family or for business - than the class they are flying in.

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

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

From this we can assume that if someone is using words like ‘delay’, ‘never’, ‘next’, that they are less likely to recommend the airline. Which is completely understandable. Noone wants to be late for vacation or wait at the airport, because the flight was delayed for a few hours or you have to get the next day flight because of some reason. From good words we can assume that if the cabin crew is friendly and provides good service, and if the seats are comfortable. the airline will more likely be recommended.

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.

We can conclude that an airline that wants to improve their customer's satisfaction should put more effort into training the cabin crew, which should be friendly and accomodating. In addition to that, the airline needs to find a good balance between number of seats on a plane and seat comfort. In case of delayed flights, they need to take proper care of passengers that are affected by it, including providing timely information on how the situation is being handled.

Although we are satisfied with the outcomes, there is still room for improvement. We could use K-fold cross validation on top of a train/test split method. The advantage of the split method is that is runs faster and is simpler than k-fold. On the other hand, the advantage of the cross-validation is a more efficient use of data since each data sample is used for training and testing. This could also be used to compute AUC scores.

Another thing that might be done is to fill out all the empty rows with average value of that specific column. For example 'type_traveller' column has only 2164 entries and we couldn't use this feature in models where we used features with more than 10000 entries. But with filling out empty rows we could. This way we would have had more data available for the modelling.