

FalconAirlines Passenger Satisfaction Case Study

by Lionel Tay Yee Dang.

TABLE OF CONTENTS

1. Project Charter	2
1.1 Problem statement	2
1.2 Business Case	2
1.3 Project Goal	2
1.4 Project scope.....	2
2. Exploratory Data Analysis	2
2.1 About the dataset.....	2
2.2 Check duplicates and null	3
2.3 Explore numerical features.....	4
2.4 Explore categorical features	4
2.5 Check correlation.....	5
2.6 Check “Age” variable	6
2.7 Investigate the cost of customer churn	7
3. Model Interpretation	8
3.1 Comparing model performance.....	8
3.2 Model interpretation (Decision Tree).....	8
3.3 Model interpretation (Logistic)	10
4. Features’ effect on customer churn cost	11
4.1 Find optimal discriminant threshold	11
4.2 Estimated reduction in customer churn cost	12
5. Conclusion	13
6. Appendix	14
6.1 Data dictionary	14
6.2 Processing pipeline Code	14

1. PROJECT CHARTER

1.1 PROBLEM STATEMENT

Airline industry is getting more competitive due to the disruption of budget airlines, which allow customers to travel to their destination at a much cheaper price though with less service quality. If the customer experience at FalconAirlines is predominantly negative, company might start losing customers and it will quickly become detrimental to the company's profit due to high capital expenditure and maintenance cost.

1.2 BUSINESS CASE

Knowing the customers' preference and feedback allows Falcon Airlines to improve their standards and services on areas that matter to the customers, and strategize the emotional selling points for their marketing campaign to stand out among their competitors, hence maintaining the company's competitiveness and customer loyalty.

1.3 PROJECT GOAL

Predict whether a passenger will be satisfied or not given the rest of the details are provided and identify which variables/features play an important role in swaying a passenger feedback towards 'satisfied'.

1.4 PROJECT SCOPE

Perform data mining on the entire customers experience journey beginning from website experience, online services, and departure to onboard experience, flight duration, and arrival.

2. EXPLORATORY DATA ANALYSIS

2.1 ABOUT THE DATASET

There are 2 datasets:

1. The 'Flight' data has information related to passengers and the performance of flights in which they travelled.
2. The 'Survey' data is the aggregated data of surveys collected post service experience.

Assumptions about the survey data:

1. **Integrity:** We assume customers are providing their honest and careful feedback for the survey.
2. **Situation & context:** We assume the customers aren't hurry to complete the survey after arriving at the destination and experiencing the service since most customers (about 69%) are business travellers as they could be busy with work. This would result in a bias outcome.
3. **Question & feedback:** We assume those survey questions are what the customers care about as the customers can't fill in any feedback other than choosing on a 1-5 scale based on those fixed questions. The outcome will

not be holistic if the company assumes these are the only questions that matter to the customers. Additional research has to be done to obtain a more accurate picture of our customers' preference and feedback.

Survey data limitation:

1. The variables/features from the survey data will not be suitable to train the model to predict the customers' satisfaction, because we will have to constantly gather the same survey data from new customers to make future prediction which doesn't make sense since we could directly ask them to indicate their satisfaction level. However, the goal of this project is to identify the important variables/features that can significantly improve customers' satisfaction. Hence, we are using the model to identify important variables/features based on their effect on the satisfaction outcome.
2. Survey is not perfect method for gathering customers' preference and feedback, hence additional research need to be done to obtain a more holistic picture.

Meta info:

After joining the flight and survey data, there are 90,917 rows and 24 columns in total. For more information about the features, descriptions, and data types (refer to Appendix 6.1 Data dictionary).

2.2 CHECK DUPLICATES AND NULL

total_duplicates <int>
0

Total duplicated counts.

variables <chr>	null_percent <dbl>
Departure_Arrival_time_convenient	9.07
Food_drink	9.00
Onboard_service	7.90
CustomerType	10.01
TypeTravel	10.00
ArrivalDelayin_Mins	0.31

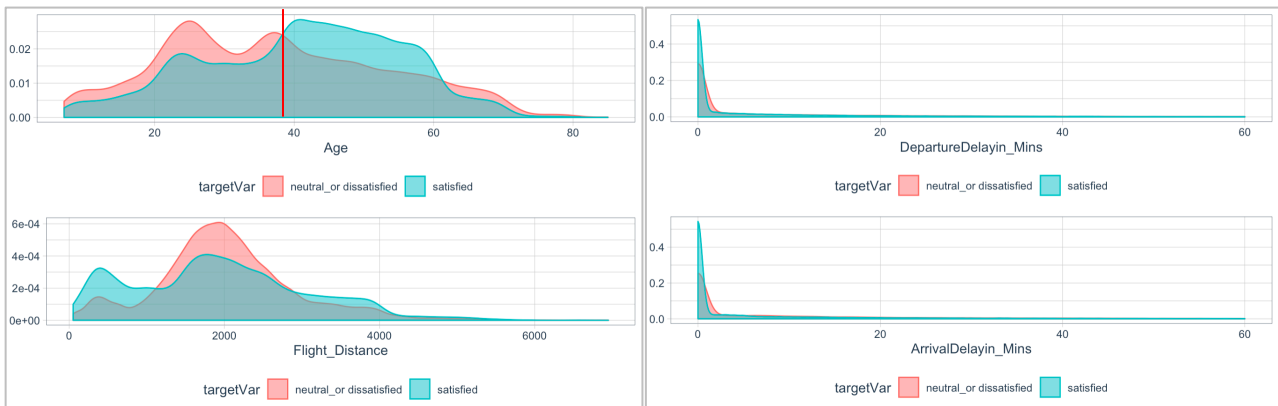
Null summary (percentage).

Findings:

1. The joined dataset contains no duplicates.
2. `CustomerType` should contain either loyal or disloyal customers. Null could probably mean the company is unable to accurately identify the customers' loyalty. I will apply Bagged Trees Imputation to fill the null.
3. `TypeTravel` should contain either business or personal travel. Null could probably mean the company is unable to identify the customers' travel motive. I will apply Bagged Trees Imputation to fill the null.
4. `ArrivalDelayin_Mins` contains 0.31% of null.
5. `Departure_Arrival_time_convenient` contains 9% of null. It should contain customer's rating. Instead of assigning an "unknown" category. I will apply Bagged Trees Imputation to fill the null so that I can transform the variables into numerical ordinal values.

- As for `Food_drink` and `Onboard_service`, null could probably mean customers did not order any food and drink or ask for any onboard service, thus no rating is given. Instead of assigning a “not applicable” category, I will apply Bagged Trees Imputation to fill the null instead so that I can transform the variables into numerical ordinal values.

2.3 EXPLORE NUMERICAL FEATURES

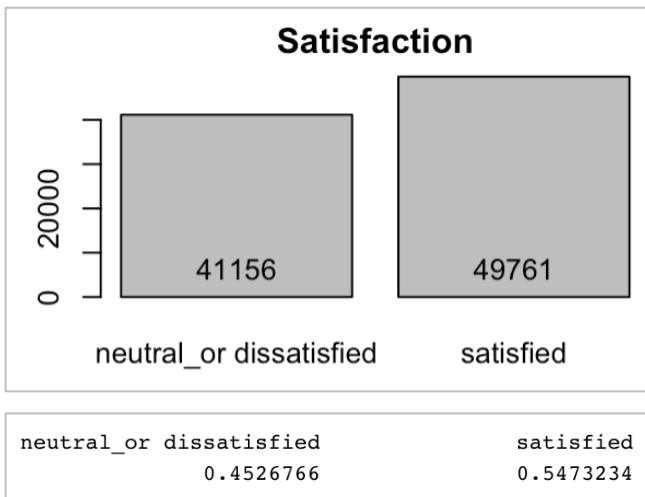


Zoomed in view for `DepartureDelayin_Mins` and `ArrivalDelayin_Mins`

Findings:

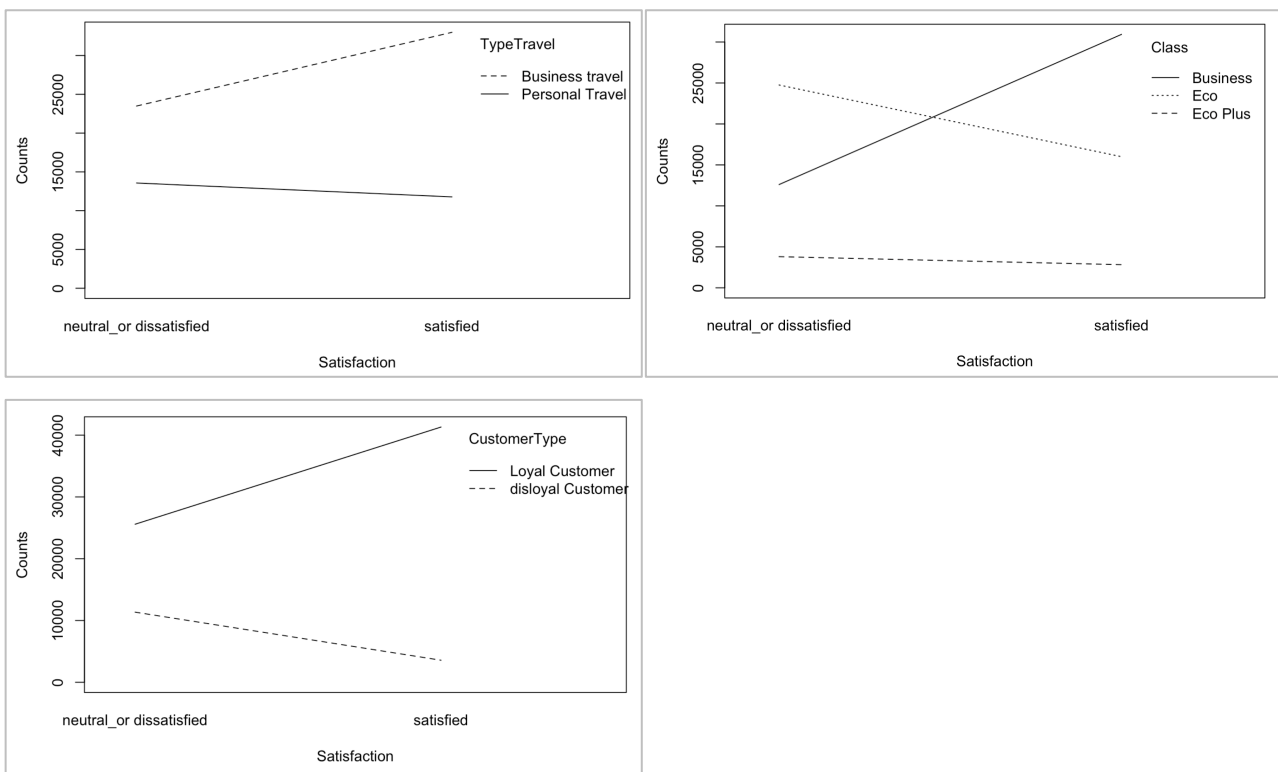
- Looking at the `Age` feature, there is a small separation between the 2 distributions. It seems that there are greater proportion of customers below age 40 feel neutral/dissatisfied than those above age 40. I will discretise the numerical value into categorical bins to improve the feature interpretability when investigating the feature weight or importance.
- Looking at the `Flight_Distance` feature, there is no significant separation between the 2 distributions. It might be a weak predictor for customers' satisfaction, hence I will remove it.
- Looking at `DepartureDelayin_Mins` and `ArrivalDelayin_Mins` features, both shows only a slight separation between the 2 distributions (red and turquoise). I will keep these variables to see how it performs.

2.4 EXPLORE CATEGORICAL FEATURES



Findings from the target variable:

- Looking at the `Satisfaction` variable, about 46% of the customers feel neutral/dissatisfied, whereas 54% feel satisfied. Hence, the dataset has a balanced class.



Findings from the variable's interaction with customer satisfaction:

- Each of the variables/features above seem to contribute a medium impact to the model predictability looking how the lines show the opposite trend or direction.

2.5 CHECK CORRELATION

	Age	Flight_Distance	DepartureDelayin_Mins	ArrivalDelayin_Mins
Age	1.00	NA	NA	NA
Flight_Distance	-0.25	1.00	NA	NA
DepartureDelayin_Mins	-0.01	0.11	1.00	NA
ArrivalDelayin_Mins	-0.01	0.11	0.97	1

Findings from the correlation between numerical variables:

1. `DepartureDelayin_Mins` and `ArrivalDelayin_Mins` have a correlation of 0.97. I will remove `ArrivalDelayin_Mins` since it contains 0.31% of null.

variable	cramer
"Inflight_entertainment"	"0.639268085016792"
"Seat_comfort"	"0.471769575989941"
"Ease_of_Onlinebooking"	"0.453261862571313"
"Online_support"	"0.431646347233728"
"Onboard_service"	"0.360286620354981"
"Online_boarding"	"0.348133711652928"
"Leg_room_service"	"0.337433279445608"
"Class"	"0.314836878500457"
"Cleanliness"	"0.304914968262666"
"CustomerType"	"0.293597832545281"
"Checkin_service"	"0.281377235539842"
"Food_drink"	"0.265839591386635"

Comparing the relationship with `Satisfaction` (target variable)

Findings from Cramer results for ordinal and categorical variables:

1. `Inflight_entertainment` seems to have very strong relationship (0.63) with `Satisfaction` (target variable), followed by `Seat_comfort`, `Ease_of_Onlinebooking`, `Online_support` and `Onboard_service`. These variables might be relatively strong predictors for customers' satisfaction.

	Online_support	Ease_of_Onlinebooking	Seat_comfort	Food_drink	Online_boarding
Online_support	1.00000000	NA	NA	NA	NA
Ease_of_Onlinebooking	0.60315929	1.00000000	NA	NA	NA
Seat_comfort	0.12232320	0.20079528	1.00000000	NA	NA
Food_drink	0.02874104	0.03458101	0.7085332	1.00000000	NA
Online_boarding	0.65091137	0.66253288	0.1318531	0.01370499	1

Findings from the Spearman correlation between ordinal variables:

1. `Ease_of_Onlinebooking`, `Online_boarding`, `Online_support` are about 60% correlated. It probably makes sense because these variables contribute to the overall user experience, hence should be taken in consideration as a whole should Falcon Airlines makes any improvement to the website or app.
2. `Food_drink` and `seat_comfort` are about 70% correlated.

I can apply variance inflation factor when training the logistic regression model to avoid multi-collinearity issue.

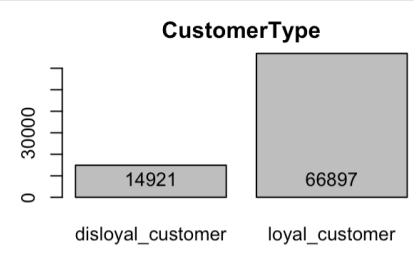
2.6 CHECK "AGE" VARIABLE

Age	Percent	Cumsum
7	0.51	0.51
8	0.60	1.11
9	0.68	1.79
10	0.62	2.40
11	0.64	3.04
12	0.61	3.65
13	0.62	4.27
14	0.66	4.94
15	0.78	5.72
16	0.90	6.61
17	0.94	7.55
18	0.93	8.48
19	0.89	9.37
20	1.49	10.86

Findings:

- Looking at the table below, there are about 8.5% of the customers who are age 18 or below. We want to focus on customers who have the purchasing/ decision power to make a purchase. Hence, I will remove respondents who are aged 18 or below as their preferences and opinions provide little value to the business, assuming that these respondents do not or have little purchasing/decision power.

2.7 INVESTIGATE THE COST OF CUSTOMER CHURN

Satisfaction <fctr>	CustomerType <fctr>	total_customer_life_time_value <chr>	CustomerType 
neutral_or dissatisfied	disloyal_customer	\$23,957,000	
neutral_or dissatisfied	loyal_customer	\$38,268,350	
neutral_or dissatisfied	NA	\$7,683,650	
satisfied	disloyal_customer	\$9,029,150	
satisfied	loyal_customer	\$93,829,650	
satisfied	NA	\$11,864,000	

Satisfaction <fctr>	CustomerType <fctr>	TypeTravel <fctr>	total_customer_life_time_value <chr>
neutral_or dissatisfied	disloyal_customer	business_travel	\$22,931,550
neutral_or dissatisfied	disloyal_customer	personal_travel	\$52,950
neutral_or dissatisfied	disloyal_customer	NA	\$972,500
neutral_or dissatisfied	loyal_customer	business_travel	\$29,359,650
neutral_or dissatisfied	loyal_customer	personal_travel	\$6,861,000
neutral_or dissatisfied	loyal_customer	NA	\$2,047,700
neutral_or dissatisfied	NA	business_travel	\$6,808,650
neutral_or dissatisfied	NA	personal_travel	\$875,000
satisfied	disloyal_customer	business_travel	\$8,676,900
satisfied	disloyal_customer	personal_travel	\$9,750

Findings:

- The life time value cost of losing the entire loyal customers who are neutral/dissatisfied is about 38.3 million. As for the "NA" customer type who are neutral or dissatisfied, assuming there are 50% of loyal customers, the life time value cost would be about 3.8 million. Total would be about 42.1 million.
- Given the entire loyal customers who are neutral/dissatisfied, the life time value cost of losing the Business Travelers is about 29.4 million compare to Personal Travellers, which is 6.9 million. While the "NA" (unknown

travel type) costs about 2 million. Given the entire “NA” customer type, assuming there are 50% of loyal customers, the total life time value cost of losing the Business and Personal Travelers would be about 3.85 million.

I will not be focusing on the life time value cost for the disloyal customers since they do not stick with FalconAirlines unless our project goal includes turning the disloyal customers into loyal via customer satisfaction.

I will later explore with the important variables to estimate the change in proportion of customer feeling satisfied and potential reduction in customer churn cost after building the model.

3. MODEL INTERPRETATION

3.1 COMPARING MODEL PERFORMANCE

The following is the results of the model performance after training and evaluating the models on the processed data (refer to Appendix 6.2 Processing pipeline to view the processing pipeline code).

model <chr>	accuracy <dbl>	f_meas <dbl>	kap <dbl>	precision <dbl>	roc_auc <dbl>	sens <dbl>
Model XGB	0.9342963	0.9263206	0.8670375	0.9214636	0.9832798	0.9312292
Model Random Forest	0.9988678	0.9987242	0.9977065	0.9982611	0.9999931	0.9991877
Model Decision Tree	0.9180333	0.9071656	0.8337917	0.9114156	0.9651462	0.9029551
Model GLM Logistic	0.8329616	0.8111264	0.6614002	0.8135725	0.9067672	0.8086950

Model performance on TRAIN set

model <chr>	accuracy <dbl>	f_meas <dbl>	kap <dbl>	precision <dbl>	roc_auc <dbl>	sens <dbl>
Model XGB	0.9300337	0.9204872	0.8580281	0.9131143	0.9815786	0.9279801
Model Random Forest	0.9523523	0.9459152	0.9033418	0.9372460	0.9916267	0.9547461
Model Decision Tree	0.9154223	0.9028450	0.8279623	0.9052242	0.9661453	0.9004783
Model GLM Logistic	0.8324904	0.8094085	0.6600066	0.8038646	0.9057824	0.8150294

Model performance on TEST set (30% rows)

Findings from the model performance table:

1. The `Random Forest` model scores the highest in all metrics. However, it has high variance due to overfitting.
2. The `XGBoost` model is the second highest, with lower variance than `Random Forest`.
3. The `Decision Tree` model is the third highest.

Since `Decision Tree` has greater model interpretability and high overall metric scores, I will go with this model instead. Nonetheless, for model interpretation, I will pair this with the `Logistic` model since the `Logistic` model is able shows us the increase or decrease in odds ratio for every 1 unit of increase for X variable.

3.2 MODEL INTERPRETATION (DECISION TREE)

Feature	Importance
"Inflight_entertainment"	"11735.61"
"Seat_comfort"	" 7535.76"
"Online_support"	" 5539.10"
"Ease_of_Onlinebooking"	" 4853.10"
"Online_boarding"	" 4023.79"
"Checkin_service"	" 2251.94"
"Food_drink"	" 1880.90"
"CustomerType_loyal_customer"	" 1140.57"
"Gate_location"	" 898.83"
"Departure_Arrival_time_convenient"	" 829.23"
"Cleanliness"	" 819.90"
"Onboard_service"	" 804.52"
"Inflightwifi_service"	" 646.38"
"Baggage_handling"	" 631.53"
"Class"	" 631.46"
"Leg_room_service"	" 613.96"
"TypeTravel_personal_travel"	" 410.03"
"Age_X_20_25_"	" 15.91"
"Age_X_65_above_"	" 10.32"
"DepartureDelayin_Mins"	" 6.97"
"Age_X_60_65_"	" 5.75"
"Age_X_35_40_"	" 4.45"
"Age_X_25_30_"	" 3.79"
"Age_X_45_50_"	" 1.71"
"Age_X_30_35_"	" 1.22"

Findings on the feature importance from `Decision Tree`:

1. The top 5 important features are `inflight_entertainment`, `Seat_comfort`, `Ease_of_Onlinebooking`, `Online_boarding`, `Online_support`. These should mostly likely be the baseline for customer satisfaction.
2. Customers at different age group do not seem to have any significant impact on customer satisfaction although customers aged between 20-25 seem to have greater impact than the rest of the age groups (perhaps due to different opinion or perspective).
3. `DepartureDelayin_Mins` also do not seem to have any significant impact on customer satisfaction, regardless of the time length. It's rather the `Depature_Arrival_time_convenient` feature that is affecting the customer satisfaction. We can find out more from the logistic model on this.

I can explore with the top N important features incrementally to investigate the business impact of the improved features (increased rating) on the proportion of customers feeling satisfied and reduction in customer churn cost. However, there is a limitation on the model interpretation. For example, in real world, even if the model is 100% accurate and we were able to increase the rating for `Inflight_entertainment` to 5 (good) or above (excellent), some customers' satisfaction might still not be affected since different customers prioritise or view the features differently. We only assume most customers would be affected by that feature.

To reduce the business risk, we can extend this research further by conducting several interviews with various segment of customers to find out more.

The following is the findings from `Decision Tree` splitting path which focuses on loyal customers only. The reason is explained in (section 2.7):

1. Inflight entertainment (≥ 5) > Customer Loyalty (≥ 5) > Class ($= 3$) > Business travel > Check-In service (≥ 4)

2. Inflight entertainment (≥ 5) > Customer Loyalty (≥ 5) > Class ($= 3$) > Business travel > Check-In service (≥ 4) > Seat comfort (≤ 4)
3. Inflight entertainment (≥ 5) > Customer Loyalty (≥ 5) > Class ($= 3$) > Business travel > Check-In service (≥ 4) > Seat comfort (≥ 5) > Gate location (≥ 5)
4. Inflight entertainment (≥ 5) > Customer Loyalty (≥ 5) > Class ($= 3$) > Personal travel > Leg room service (≥ 5)
5. Inflight entertainment (≥ 5) > Customer Loyalty (≥ 5) > Class (≤ 2) > Seat comfort ($= 6$) > Check-In service (≥ 4)
6. Inflight entertainment (≥ 5) > Customer Loyalty (≥ 5) > Class (≤ 2) > Seat comfort (≤ 5) > Departure arrival time convenient (≤ 5) > Food drink (≤ 4)
7. Inflight entertainment (≥ 5) > Customer Loyalty (≥ 5) > Class (≤ 2) > Seat comfort (≤ 5) > Departure arrival time convenient (≤ 5) > Food drink (≥ 4) > Gate location (≥ 5)
8. Inflight entertainment (≥ 5) > Customer Loyalty (≥ 5) > Class (≤ 2) > Seat comfort (≤ 5) > Departure arrival time convenient (≥ 5) > Personal travel
9. Inflight entertainment (≥ 5) > Customer Loyalty (≥ 5) > Class (≤ 2) > Seat comfort (≤ 5) > Departure arrival time convenient (≥ 5) > Business travel > Seat comfort (≥ 5)

3.3 MODEL INTERPRETATION (LOGISTIC)

Feature	Log Odds	Odds Ratio
"(Intercept)"	"-11.5219"	"0.000"
"CustomerType_loyal_customer"	"2.0452"	"7.731"
"TypeTravel_personal_travel"	"-0.7768"	"0.460"
"Inflight_entertainment"	"0.7158"	"2.046"
"Age_X_20_25_"	"0.6252"	"1.869"
"Class"	"0.3470"	"1.415"
"Age_X_65_above_"	"-0.3296"	"0.719"
"Onboard_service"	"0.3095"	"1.363"
"Checkin_service"	"0.2986"	"1.348"
"Seat_comfort"	"0.2916"	"1.339"
"Leg_room_service"	"0.2462"	"1.279"
"Ease_of_Onlinebooking"	"0.2250"	"1.252"
"Departure_Arrival_time_convenient"	"-0.2233"	"0.800"
"Food_drink"	"-0.1926"	"0.825"
"DepartureDelayin_Mins"	"-0.1810"	"0.834"
"Online_boarding"	"0.1587"	"1.172"
"Gate_location"	"0.1285"	"1.137"
"Online_support"	"0.1010"	"1.106"
"Baggage_handling"	"0.0934"	"1.098"
"Cleanliness"	"0.0917"	"1.096"
"Inflightwifi_service"	"-0.0651"	"0.937"

Findings on the feature weight from `Logistic`:

1. The feature importance order from the `Decision Tree` is mostly different from the `Logistic` model, although some do follow the similar order.
2. `Inflight_entertainment` increases the odds ratio (probability of feeling satisfied vs neutral/dissatisfied) by 2 times for every 1 unit of increase in rating. It is a top performing feature similar to the feature importance result from the `Decision Tree` model.
3. `Class` increases the odds ratio by 40% times for every 1 unit of increase in class level (economic > economic plus > business).

4. The age group between 20–25 seems to increase the odds ratio by 1.8 times although the feature importance result from the `Decision Tree` model does not show any significant impact on customer satisfaction.
5. `Onboard_service`, `Checkin_service`, `Seat_comfort` increases the odds ratio by roughly 35% for every 1 unit of increase in rating.
6. The rest contributes about 10–25% of increment in odds ratio.

Findings on significant negative performing feature:

1. Personal Travelers seems decrease the odds ratio by 54%. Falcon Airlines should give extra attention to Personal Travelers and get more detailed feedback from them. Although they only occupy about 30% of the current customers data and has lower customer churn cost (7.7 million) than Business Travelers (29.3 million) (refer to section 2.7), however, losing half of these customers may also be detrimental to the airline business since most airlines tend to have higher expenditure and capital spending (meaning low profit margin).
2. Next is `Departure_Arrival_time_convenient`, which decreases the odds ratio by 20% for every 1 unit of increase in rating, which is weird since customers who rate this higher should not make them feel less satisfied. The same goes for `Food_drink` variable. There could be a data error, thus should investigate further. Or, it could also be due to the customer priority. For example, departure/arrival time convenient and food might not affect customer's satisfaction although they rate it low as other features matter more to them.

4. FEATURES' EFFECT ON CUSTOMER CHURN COST

4.1 FIND OPTIMAL DISCRIMINANT THRESHOLD

Satisfaction <fctr>	CustomerType_loyal_customer <dbl>	TypeTravel_personal_travel <dbl>	total_customer_life_time_value <chr>
0	0	0	\$20,283,150
0	0	1	\$43,000
0	1	0	\$26,217,300
0	1	1	\$5,949,150
1	0	0	\$7,419,450
1	0	1	\$6,500
1	1	0	\$73,082,100
1	1	1	\$5,140,050

Actual customer life time value on TRAIN set.

.pred_class <fctr>	CustomerType_loyal_customer <dbl>	TypeTravel_personal_travel <dbl>	total_customer_life_time_value <chr>
0	0	0	\$23,351,850
0	0	1	\$44,300
0	1	0	\$26,604,450
0	1	1	\$6,496,500
1	0	0	\$4,350,750
1	0	1	\$5,200
1	1	0	\$72,694,950
1	1	1	\$4,592,700

Estimated customer life time value on TRAIN set after adjusting the threshold to 0.75

To avoid underestimating the customer churn cost for loyal business and personal travellers who are neutral/dissatisfied, the discriminant threshold should at least be around 0.75.

4.2 ESTIMATED REDUCTION IN CUSTOMER CHURN COST

I will first explore the following top 5 important features to see the effect by increasing the rating to 5 (good) or above (excellent). Then I will estimate the customer life time value after increasing the rating for these 5 features and making model prediction:

1. Inflight entertainment.
2. Seat comfort.
3. Online support.
4. Ease of online booking.
5. Online boarding.

Satisfaction <fctr>	CustomerType_loyal_customer <dbl>	TypeTravel_personal_travel <dbl>	total_customer_life_time_value <chr>
0	0	0	\$2,810,450
0	0	1	\$12,550
0	1	0	\$3,643,400
0	1	1	\$2,550,400
1	0	0	\$1,088,650
1	0	1	\$4,550
1	1	0	\$10,460,350
1	1	1	\$2,280,450

Customer churn cost on TEST data (before improving the rating)

.pred_class <fctr>	CustomerType_loyal_customer <dbl>	TypeTravel_personal_travel <dbl>	total_customer_life_time_value <chr>
0	0	0	\$3,468,550
0	0	1	\$14,500
0	1	0	\$1,332,050
0	1	1	\$1,320,450
1	0	0	\$430,550
1	0	1	\$2,600
1	1	0	\$12,771,700
1	1	1	\$3,510,400

Customer churn cost on TEST data (after improving the rating)

Findings after increasing the rating, making prediction, and calculating the value on TEST data (30% rows):

1. After calculating the difference between the before and after, the company is able to save 3.54 million (6.19 – 2.65 million) roughly on loyal Business and Personal Travellers who are neutral/dissatisfied, which is around 57% of reduction just by focusing on improving the top 5 important features.

	percentile	prob
[1,]	"0%"	"0.0000"
[2,]	"2.5%"	"0.0000"
[3,]	"5%"	"0.2941"
[4,]	"7.5%"	"0.3012"
[5,]	"10%"	"0.3012"
[6,]	"12.5%"	"0.3012"
[7,]	"15%"	"0.3012"
[8,]	"17.5%"	"0.3237"
[9,]	"20%"	"0.3237"
[10,]	"22.5%"	"0.6031"
[11,]	"25%"	"0.6031"
[12,]	"27.5%"	"0.6031"
[13,]	"30%"	"0.7554"

Percentile of predicted probability

neutral_or dissatisfied	satisfied
0.4526766	0.5473234

Original class proportion

Findings on the predicted probability by Decision Tree after increasing the rating:

1. The proportion of the predicted neutral/dissatisfied customers is reduced from 45% to between 22.5–30% (between 0.5 to 0.75 discriminant threshold), which is 15–22.5% of difference just by focusing on improving the top 5 important features.

CustomerType_loyal_customer	Satisfaction	Counts
<dbl>	<fctr>	<int>
0	0	3426
0	1	1110
1	0	7446
1	1	12930

Class distribution (before improving the rating)

CustomerType_loyal_customer	.pred_class	Counts
<dbl>	<fctr>	<int>
0	0	4101
0	1	435
1	0	3160
1	1	17216

Class distribution (after improving the rating)

Findings:

1. The table shows around 57.6% of reduction $((7446 - 3160) / 7446)$ for neutral/dissatisfied loyal customers just by focusing on improving the top 5 important features.

5. CONCLUSION

The following recommendations focus on loyal customers who are neutral/dissatisfied:

1. Extend the user research to interviews with different segment of customers after performing the customer segmentation, especially with personal travellers to find out more on customer preference. We can also inquire them about the top N important features to understand why they matters to the customers. We can also find out what makes a customer loyal to our FalconAirlines brand so that we can also target the disloyal customers who are neutral/dissatisfied.
2. Extend the user research to value mapping to map out and analyse the customer touchpoint and user journey and to better understand the detail of the customer journey and the interaction on each touchpoint.
3. Investigate the `Departure_Arrival_time_convenient` and `Food_drink` data as there could be a data error (refer to section 3.3 on negative performing feature – no. 2).

4. Analyse the current business capabilities and feature feasibility for the top N important features before moving to the business transformation planning. We can then estimate the potential reduction in customer churn cost for features that the business is capable to plan and execute for that time period.
5. We might want to pay extra attention to loyal business travellers since the customer churn cost is higher as they travel more frequently than the personal travellers. Plus, about 69% of our customers are business travellers based on the current data.

6. APPENDIX

6.1 DATA DICTIONARY

Columns	Description	Dtypes
CustomerId	Reference ID.	–
Satisfaction	Indicating whether customers are satisfied or not.	Factor
Gate_location	How convenient the customers feel about the gate location.	Ordinal
Seat_comfort	How comfortable the customers feel about the seat.	Ordinal
Departure_Arrival_time_convenient	How convenient the customers feel about the departure and arrival time.	Ordinal
Food_drink	Quality of the food and drink.	Ordinal
Inflightwifi_service	Quality of the inflight wifi service.	Ordinal
Inflight_entertainment	Quality of the inflight entertainment. Could be games or others.	Ordinal
Online_support	Quality of the online support.	Ordinal
Ease_of_Onlinebooking	Quality of the online booking experience.	Ordinal
Onboard_service	Quality of the inflight service.	Ordinal
Leg_room_service	Leg room space.	Ordinal
Baggage_handling	Quality of the baggage handling and collecting process.	Ordinal
Checkin_service	Quality of the check-in service.	Ordinal
Cleanliness	Cleanliness.	Ordinal
Online_boarding	Quality of the online check-in service.	Ordinal
Gender	Gender.	Factor
CustomerType	Loyal, disloyal, or unknown customers.	Factor
TypeTravel	Business travel, personal travel, or others.	Factor
Class	Economic, economic plus, or business class.	Ordinal
Age	Age.	Num
Flight_Distance	Flight distance (km).	Num
DepartureDelayin_Mins	Delay in departure time (minutes).	Num
ArrivalDelayin_Mins	Delay in arrival time (minutes).	Num

6.2 PROCESSING PIPELINE CODE

```

# -----
# Prepare the variables.
# -----

# To indicate which variables will be used for Bagging Imputation.
impute_with_vars = df_falconAirlines %>%
  select(-CustomerId, -Satisfaction, -Gender, -ArrivalDelayin_Mins, -Flight_Distance) %>%
  colnames()

# To indicate which variables should be processed.
remove_vars = c('CustomerId', 'ArrivalDelayin_Mins', 'Flight_Distance', 'Gender')
missing_vars = c('CustomerType', 'TypeTravel', 'Departure_Arrival_time_convenient', 'Onboard_service', 'Food_drink')
binning_vars = c('Age')
dummy_vars = c('CustomerType', 'TypeTravel', 'Age')
dist_transform_vars = c('DepartureDelayin_Mins')
normalise_vars = c('DepartureDelayin_Mins')

# -----
# Construct processing pipeline.
# -----

recipe_obj = recipe(Satisfaction ~ ., data=df_train) %>%
  step_rm(remove_vars) %>%
  step_filter(Age > 18) %>%
  step_bagimpute(missing_vars, impute_with=impute_with_vars, seed_val=5) %>%
  step_mutate_at(binning_vars, fn=binning) %>%
  step_dummy(dummy_vars) %>%
  step_rename_at(all_predictors(), fn= ~ str_replace_all(., '\\\\.', '_')) %>%
  step_mutate_at(all_predictors(), fn=as.numeric) %>%
  step_YeoJohnson(dist_transform_vars) %>%
  step_normalize(normalise_vars) %>%
  prep()

```