University of Warsaw

Faculty of Economic Sciences

Jakub Gazda — Student no.: 419 272

Aleksander Wieliński — Student no.: 420 272

„Analysis of airline customer satisfaction and key satisfaction drivers using a binary dependent variable model.”

Advanced Econometrics project

Under the guidance of:

Dr. Rafał Woźniak

Warszawa, czerwiec 2023

Abstract

Within this paper, we employed an econometric model to analyze a satisfaction survey database and identify the key factors influencing travel satisfaction. The dataset was cleaned, variables were prepared, and relevant interactions were created. Model comparison using information criteria and R-squared guided our selection of the probit model for further analysis. After eliminating insignificant variables and conducting diagnostic tests, we examined our hypotheses based on the marginal effects of the variables. Our findings confirmed all of our hypotheses, suggesting important insights for enhancing the aviation industry's focus on service quality and creating enjoyable travel experiences.

Introduction

Airline business is among one of the most crucial within travel industries in the world. With many people utilizing this way of transport, especially now after COVID, it is important to understand what are the key factors to customer satisfaction, in order to better optimize flights. This can also lead to positive environmental changes, as using such analyses, it was discovered that less fuel can be wasted, since customers don’t mind flying for longer times. Among the many reasons for passenger’s satisfaction when flying, recently there was a great focus on environmental friendliness and the quality of services provided by the aviation companies, but this will be explained deeper in the next chapters. However, this change of focus is important to the understanding of the air traveling market and the areas companies can improve to gather more clients.

Based on the literature provided in the next chapter and our intuit we decided to verify three hypotheses regarding the satisfaction of air travel:

H1: Distance of travel has no influence on customer satisfaction.

H2: Loyal customers are more forgiving and will be satisfied more often than the disloyal customers.

H3: Meeting customers' needs and requirements, and exceeding their expectations in quality of provided services lead to a higher probability of satisfaction.

Chapter I: Literature review

As with every service industry, the goal of the companies is to exceed customer expectations with their service ultimately outshining the competition. But the aviation sector is a highly saturated market where customers pick and choose from many different options, that is why passenger gratification is so crucial and essential for the companies providing these services[1].

There was a major shift of focus on cost saving towards quality improvements as the industry became more and more challenging, when the chase for the fastest travel time faded out and converted into the competition of the best and most friendly processes involving a passenger. In addition to that customers nowadays seem to be extremely quality-sensitive highlighting the stepback from being the fastest airline towards being the most reliable, easy to use, customer friendly services[2].

When it comes to service quality itself, there are many publications discovering different specific services having an influence on satisfaction level. Each research included a different approach to the problem, however they often indirectly proposed two groups of main categories:

- *pre-flight services;*
- *in-flight services;*

The first group might be a little confusing as how can a service be judged before the flight happens? As many articles pointed out, pre-flight services include not only all the booking-related processes but everything until a departure. There were positive correlations discovered between satisfaction and those services, and some even argue that they had the most impact on the satisfaction levels[3]. There are also some factors considered an activity conducted both before flight and after, like handling the baggage, however they are oftentimes bundled up together with pre-flight services.

The second group refers to a much larger cluster of services that are provided during the flight itself. Different sources tackle the problem by looking at slightly different metrics - some measure based upon the quality of food, some chair comfort, some interaction with staff

[1] Hayadi, B. H., Kim, J. M., Hulliyah, K., & Sukmana, H. T. (2021). Predicting Airline Passenger Satisfaction with Classification Algorithms. International Journal of Informatics and Information System, 4(1), p. 82.
[2] Namukasa, J. (2013). *The influence of airline service quality on passenger satisfaction and loyalty. The TQM Journal*, 25(5), p. 520-522.
[3] Ibidem, p. 527.

members, but there were metrics that used to match, like tidiness of the plane or provided entertainment - real concepts measurable on the plane by touch, sight and general feel[4].

But other than quality of services there is also a suspicion of another factor having its effect on the satisfaction level. It has been discovered that customer loyalty and satisfaction are in fact correlated with each other. As briefly mentioned before, many previous findings also hinted at the fact of loyal customers being a great influence on satisfaction level and every action trying to improve the satisfaction should be aimed at them, as they are the most high-value being less prone to turn to competitors[5]. The relationship between disloyalty/loyalty and services has also been thoroughly studied, stating the importance of increasing the above mentioned quality of *in-flight services* indirectly influencing customer satisfaction through passenger's connection to the brand.

Overall quality has been identified as a major contributor to customer satisfaction, especially linking the most profitable ones with the high tendency to be attracted to superior quality measures[6], but based on the provided literature we also suspect other variables being connected with the satisfaction or dissatisfaction of the passenger. In this paper based on available data, provided literature and intuition we decided to verify the following hypothesis regarding the satisfaction levels:

H1: Distance of travel has no influence on customer satisfaction.

H2: Loyal customers are more forgiving and will be satisfied more often than the disloyal customers.

H3: Meeting customers' needs and requirements, and exceeding their expectations in quality of provided services lead to a higher probability of satisfaction.

Chapter II: Preparation

Data

The dataset which upon the model will be conducted was taken from the Kaggle website (Airlines Customer satisfaction | Kaggle) and contain basic information about quality of

[4] Hussain, R., Al Nasser, A., & Hussain, Y. K. (2015). *Service quality and customer satisfaction of a UAE-based airline: An empirical investigation. Journal of Air Transport Management*, 42, p. 168.

[5] Namukasa, J. (2013). *The influence of airline service quality on passenger satisfaction and loyalty. The TQM Journal*, 25(5), p. 523-529.

[6] Jiang, H., & Zhang, Y. (2016). *An investigation of service quality, customer satisfaction and loyalty in China's airline market. Journal of Air Transport Management*, 57, p. 87.

service, flight statistics, passenger statistics and overall satisfaction displayed in almost 130k observations using 23 variables total. Based on the provided literature and intuition we decided to include all of them in our model as a good starting point:

- **Satisfaction** - Categorical variable with two values: "satisfied" and "dissatisfied" indicating customer satisfaction level;
- **Gender** - Gender of the customer, can be either "Male" or "Female";
- **Customer.type** - Category of the customer based on their type, with two possible values: "Loyal Customer" or "Disloyal Customer";
- **Age** - Numerical variable representing the customer's age in years;
- **Type.of.travel** - Purpose of the customer's travel, categorical variable with two values: "Business travel" or "Personal travel";
- **Class** - Class of service the customer availed, categorical variable with three values: "Eco", "Eco Plus", or "Business";
- **Flight.distance** - Distance of the flight in miles, numerical variable;
- **Seat.comfort** - Customer's satisfaction with seat comfort, numerical variable measured on a scale from 0 to 5, where higher values indicate greater satisfaction;
- **Departure/Arrival.time.convenient** - Customer's satisfaction with the convenience of departure and arrival times, numerical variable measured on a scale from 0 to 5, where higher values indicate greater satisfaction;
- **Food.and.drink** - Customer's satisfaction with the food and drink provided during the flight, numerical variable measured on a scale from 0 to 5, where higher values indicate greater satisfaction;
- **Gate.location** - Customer's satisfaction with the gate location, numerical variable measured on a scale from 0 to 5, where higher values indicate greater satisfaction;
- **Inflight.wifi.service** - Customer's satisfaction with the inflight wifi service, numerical variable measured on a scale from 0 to 5, where higher values indicate greater satisfaction;
- **Inflight.entertainment** - Customer's satisfaction with the inflight entertainment options, numerical variable measured on a scale from 0 to 5, where higher values indicate greater satisfaction;
- **Online.support** - Customer's satisfaction with the online customer support, numerical variable measured on a scale from 0 to 5, where higher values indicate greater satisfaction;

- **Ease.of.online.booking** - Customer's satisfaction with the ease of online booking, numerical variable measured on a scale from 0 to 5, where higher values indicate greater satisfaction;
- **Onboard.service** - Customer's satisfaction with the on-board service provided by the airline, numerical variable measured on a scale from 0 to 5, where higher values indicate greater satisfaction;
- **Leg.room.service** - Customer's satisfaction with the leg room provided during the flight, numerical variable measured on a scale from 0 to 5, where higher values indicate greater satisfaction;
- **Baggage.handling** - Customer's satisfaction with the airline's baggage handling, numerical variable measured on a scale from 0 to 5, where higher values indicate greater satisfaction.

Before we proceed it is important to mention that the dataset was split due to the size of the model, resources it took to process and calculate the model as well as marginal effects, and memory restrictions. It was decided to use a fraction of the model acting as a demonstrative split indicating directions for further studies. The final but smaller dataset contained 650 observations but retained all 23 variables.

Clearing

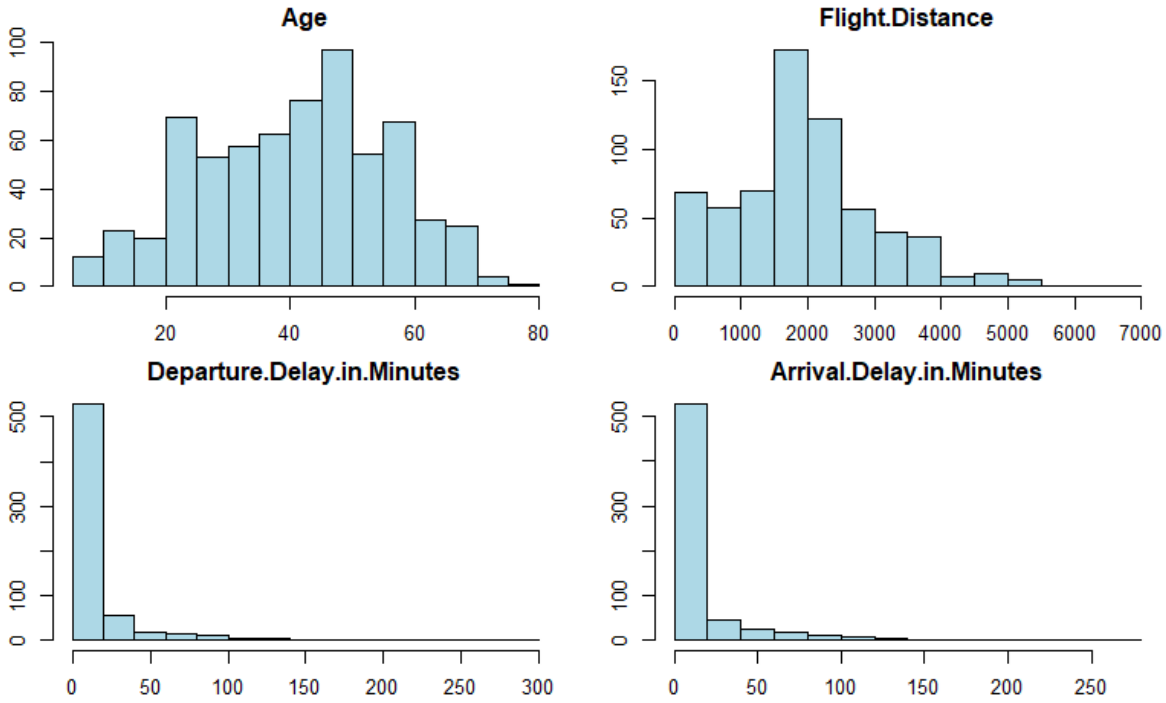
The first actual step to the modeling is to preprocess the data, starting with missing values. Our data contained 3 observations with missing values, and due to the small number of them we decided to delete them without further investigation.

This was followed by mapping categorical variables with digit values, so that the categories would be represented numerically. This was conducted upon target variable - Satisfaction - but also on Gender, Customer.type, Type.of.travel and Class, with the latter being the only non-binary variable as it was an ordinal variable. The key takeaway from that transformation is that the dependent variable is now represented by binary values - 1 for satisfied answers and 0 for dissatisfied.

Before making even more changes, let’s investigate distributions of the variables to further understand the dataset. On Figure 1 there are distributions of all the continuous variables in the dataset. From a quick glance we can assume variable Age having a normal distribution, and Flight.distance having one as well, however with a presence of outliers skewing the distribution

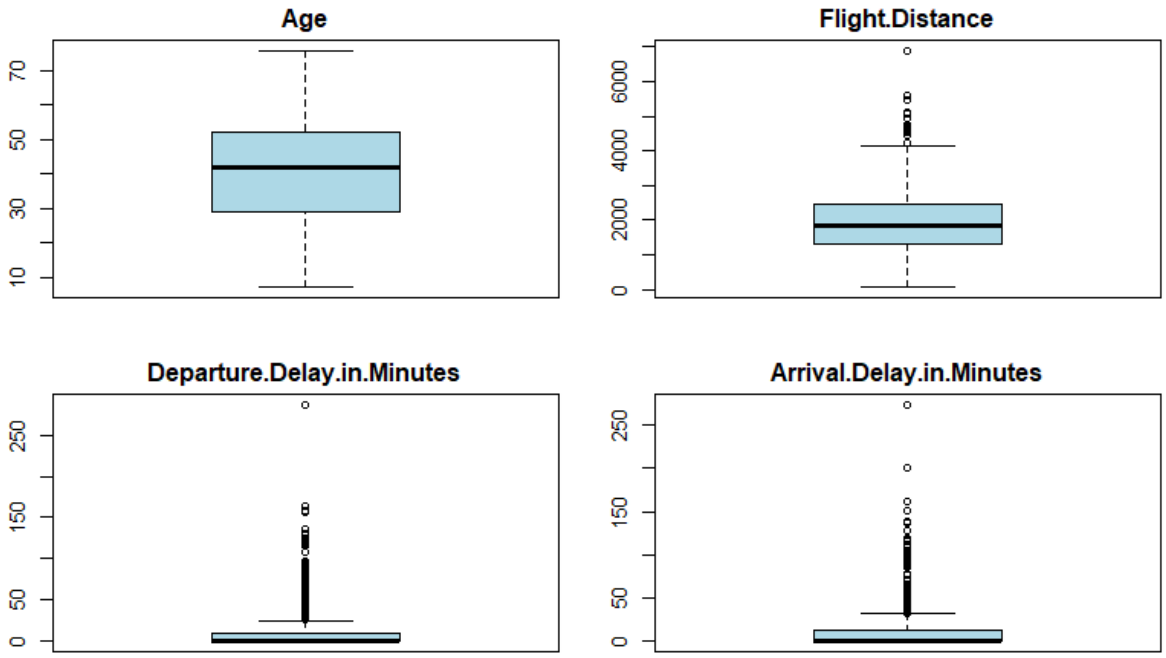
to the right. The other two variables resemble geometric distributions with a strong suspicion of outliers.

Figure 1. Histograms of the continuous variables in our dataset.



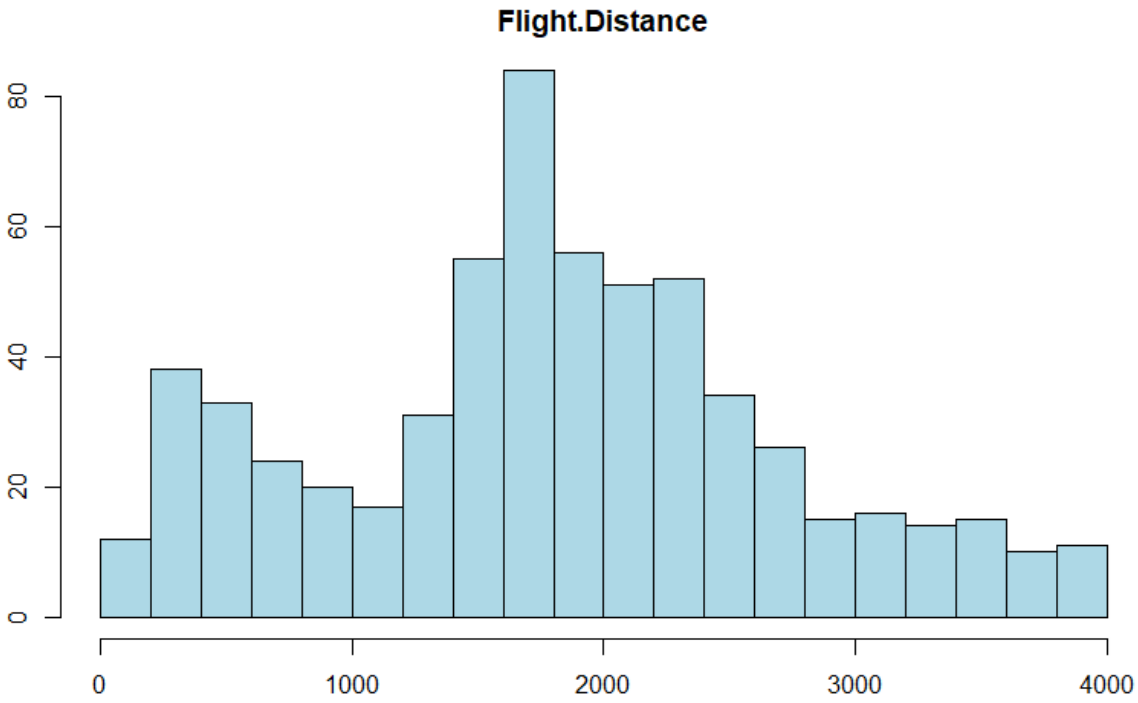
This is further explored using the box-plots on Figure 2 where there is a noticeable amount of outliers present, especially in the Departure/Arrival.in.minutes. However the Flight.distance variable also contains a fair amount of outliers. The pleasant surprise is found in the Age variable which doesn't contain any outliers and forms a normal distribution.

Figure 2. Box-plots of the continuous variables in the dataset.



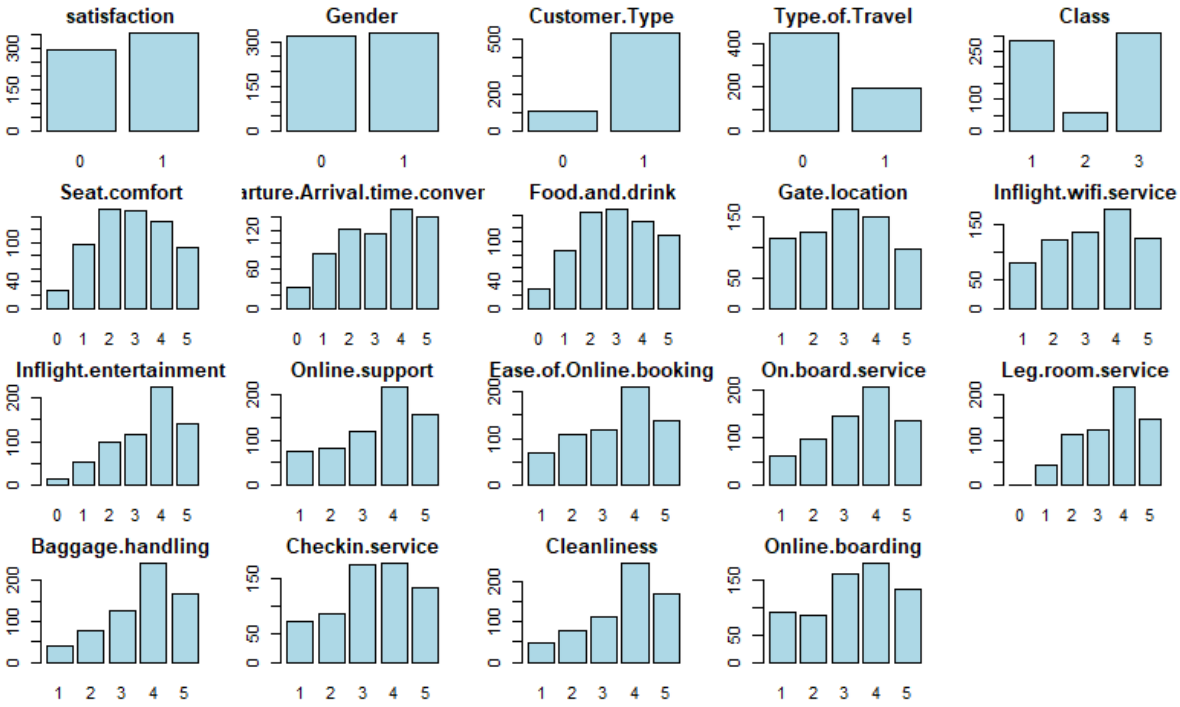
We deemed Departure/Arrival.in.minutes hopeless, and decided to delete outliers from the Flight.distance variable, setting a cutoff point at the 0.95 quantile. Figure 3 displays the distribution after cutting 5% of the variables, and as we can see the distribution is no longer skewed to the right, instead it looks similar to the normal distribution.

Figure 3. Histogram of the Flight.distance variable after removing 5% of variables.



The other variables, presented on Figure 4, tell a different kind of story about a balanced dataset, at least from the perspective of the target variable, displaying a similar volume of satisfaction and dissatisfaction. There are of course few variables, especially binary and ordinal, that present a discrepancy of one variable that is dominated by the other, such as in the Customer.type or Class. Here we can also see bar plots of all the variables taken from a survey having values from 0 to 5. They were bundled up here with the rest of the categorical variables as they can be interpreted as such as each score represents a different level of satisfaction from given service, however moving forward it was decided that they will be treated as discrete variables. The decision was taken upon trying an approach of encoding them dichotomously, however the loss of information from that transformation was too great for the model to have a proper form, thus the reason to abandon that idea.

Figure 4. Bar plots of binary and categorical variables in the dataset.



Overall base variables were mostly cleaned up and preprocessed to a proper form so they could be used in the modeling. Other than that missing values and outliers were removed to provide a better fit to the model, concluding the preparation of the variables from the dataset.

Exploratory Data Analysis and Feature Engineering

Having the data preprocessed we can go to the next step, which is a deep look at the ongoing connections between the features and the dependent variable. The correlation map shown on Figure 5 is very informative when it comes to discovering hidden infrastructure of relations in our dataset. There is no surprise that some variables with satisfaction survey scores are correlated with each other, as they came from the same category of services. Some variables were also suspected to have a joint influence on the target variable by creating an interaction. Based on the intuition, provided literature and correlations between variables we created and implemented to the model the following interactions:

- Seat.comfort*Food.and.drink;
- Departure.Arrival.time.convenient*Gate.location;
- Food.and.drink*Gate.location;
- On.board.service*Cleanliness;
- Inflight.entertainment*Online.boarding;

Each generated interaction makes sense for those who used the aviation services at least once. Seat comfort and beverages create a combination of quality needs that can influence the comfort of the flight, just like on board services and tidiness of the plane. Interactions with the location of the gate can be also easily explained as both food stands and the process of preparing the plane is related to the position and ease of access to the gate. The last interaction being entertainment on the flight and online boarding was taken from the assumption that the companies technologically advanced enough to provide online services will be also capable of providing digital entertainment to the passengers.

Figure 5. Heatmap of correlations between all variables in the dataset.

By looking at the target variable - Satisfaction - we can see several variables positively correlated, some not correlated at all and few weak negative correlations, with the highest being Inflight.entertainment and the lowest Gender. After preprocessing the dataset and choosing the interactions It was decided to include all the variables in the base model as each of them seem important to the analysis based on the provided literature and the intuition. Now, having the variables ready we can proceed further with the analysis towards choosing the most appropriate method of actually gathering insight about the influence on the target variable - the Binary Dependent Variable Models.

Chapter III: Methodology

Comparing models

To begin with modeling and evaluation, a crucial decision has to be made between the choice of logit and probit model. Since the database contains mostly normally distributed variables it is appropriate to choose the probit model, however the further comparison is required using information criteria. We chose Akaike Information Criteria as our baseline and based on the value of 431.8066 for the probit model, ang logit's AIC of 429.5716 we are supposed to go with the logit, as it has a lower AIC value. There is however a condition of using such criteria reliably, as they only provide information when the dataset is noticeably imbalanced[7] - which is not our case as shown on the plots in the previous chapter. Discrediting the AIC criteria information approach leaves us with the initial method of choosing the appropriate model, thus we are going forward with probit as also supported by the BIC criteria, which yields LPM worse.

Figure 6. Information Criterion of the chosen models

Model	AIC	BIC
LPM (with robust matrix)	429.9717	558.1515
Logit	429.5716	553.3314
Probit	431.8066	555.5665

Looking at the R-squared statistics between logit and probit we can see each of them being ever so slightly better than the other, so there is no clear winner supporting our choice. However it's important to mention that no matter the outcomes of many R-squared values, the goal of the model is not to maximize the percentage of explained variance of the model, but to reflect the reality the best.

[7] Chen, G., & Tsurumi, H. (2010*). Probit and Logit Model Selection. Communications in Statistics - Theory and Methods*, 40(1), p. 174.

Figure 7. Information Criterion of the chosen models

Model	R-squared	McFadden	McKelveyZavoina
Logit	0.4417318	0.5582682	0.7620465
Probit	0.4443746	0.5556254	0.7627257

General-to-specific

After settling on the probit model in the analysis, a necessary selection of key variables had to be made. Figure 8 shows all of the variables with their interactions, their coefficients and statistical significance. In order to conclude which variables are statistically insignificant and can be removed from the model safely, we have conducted a general to specific procedure, with use of the likelihood ratio test.

Figure 8. Summary of a base probit model with all variables compared to the final probit.

	Dependent variable:	
	satisfaction	
	(1)	(2)
Gender	-0.569***	-0.667***
	(0.158)	(0.149)
Customer.Type	1.253***	1.208***
	(0.265)	(0.239)

Age	-0.009	
	(0.005)	
Type.of.Travel	-0.676***	-0.683***
	(0.237)	(0.184)
Class	0.076	
	(0.105)	
Flight.Distance	-0.0001	
	(0.0001)	
Seat.comfort	-0.979***	-0.942***
	(0.146)	(0.139)
Food.and.drink	-0.488***	-0.458***
	(0.180)	(0.171)
Departure.Arrival.time.convenient	-0.636***	-0.616***
	(0.153)	(0.146)
Gate.location	-0.116	-0.128
	(0.186)	(0.179)
Inflight.wifi.service	0.001	
	(0.083)	
Inflight.entertainment	0.032	0.012

	(0.178)	(0.172)
Online.support	0.132	
	(0.081)	
Ease.of.Online.booking	0.244**	0.343***
	(0.113)	(0.093)
On.board.service	-0.361*	-0.298
	(0.209)	(0.194)
Leg.room.service	0.150*	
	(0.077)	
Baggage.handling	0.083	
	(0.092)	
Checkin.service	0.128**	0.164***
	(0.065)	(0.062)
Cleanliness	-0.408**	-0.399**
	(0.183)	(0.174)
Online.boarding	-0.282	-0.307*
	(0.190)	(0.177)
Arrival.Delay.in.Minutes	0.018	
	(0.021)	

Departure.Delay.in.Minutes	-0.032	
	(0.021)	
Seat.comfort:Food.and.drink	0.411***	0.401***
	(0.047)	(0.045)
Departure.Arrival.time.convenient:Gate.location	0.165***	0.162***
	(0.050)	(0.047)
Food.and.drink:Gate.location	-0.201***	-0.199***
	(0.057)	(0.055)
On.board.service:Cleanliness	0.121**	0.120**
	(0.055)	(0.051)
Inflight.entertainment:Online.boarding	0.115**	0.123**
	(0.050)	(0.048)
Constant	0.981	1.191
	(0.994)	(0.865)
Observations	614	614
Log Likelihood	-187.903	-194.287
Akaike Inf. Crit.	431.807	426.575

Note: $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

As can be seen in the above table, we have identified 13 statistically insignificant variables. At first, a null probit model was created, assuming a constant of 1 as the only variable. The likelihood ratio test was then performed on the null model as well as a default starting probit model. With the Chi squared statistic of 469.89 and p-value below 2.2e-16 (near zero) we can reject the null hypothesis that the nested model (null probit) fits the data better, in favor of the alternative hypothesis, that the full model fits the data significantly better. We have then proceeded to test the nested models, to evaluate the significance of variable removal, going in order of the highest p-value. It is worth mentioning that certain variables, even though statistically insignificant, were left in the model. This was caused by the fact that they were part of an interaction between variables, therefore making it a crucial part of the model.

Tests and diagnostics

Once done with the general to specific procedure, it was necessary to appropriately diagnose and test the final model. Firstly, the Link test is performed. The test verifies, whether the assumed model is not misspecified, and whether additional independent variables can be found. As can be seen on the below Figure 9, we can conclude that yhat is statistically significant with a Z value of 13.594 and corresponding p-value near zero. Yhat2 is statistically insignificant, with a Z value of 0.462 and a corresponding p-value of 0.644. This is evidence of the fact that the model is indeed well specified, and no additional independent variables can be found.

Figure 9. Link test of our final probit.

	Dependent variable:
	y
yhat	0.552***
	(0.041)
yhat2	0.008
	(0.017)
Constant	-0.024
	(0.088)
Observations	614
Log Likelihood	-196.645
Akaike Inf. Crit.	399.290
Note:	*p<0.1; **p<0.05; ***p<0.01

As a next step, we are performing the Hosmer-Lemeshow goodness of fit test. With a p-value of 0.218 and X squared statistic of 10.721, we can conclude that there is no evidence of the predicted probabilities deviating from observed ones. Both linktest and Hosmer-Lemeshow seem to confirm the proper form of our model.

Odds ratio

Because a probit model is under analysis, and not a logit model, we omit the analysis of odds ratio. While it is possible somehow for probit models, it proves vastly impractical, since the odds vary for each variable value, complicating the interpretation.

Chapter IV: Results

Figure 10. All the models throughout our analysis.

	Dependent variable:				
	satisfaction				
	LPM	*Robust LPM*	*Logit*	*Base Probit and Final Probit*	
	(1)	(2)	(3)	(4)	(5)
Gender	-0.119***	-0.119***	-1.016***	-0.569***	-0.667***
	(0.029)	(0.029)	(0.288)	(0.158)	(0.149)
Customer.Type	0.263***	0.263***	2.317***	1.253***	1.208***
	(0.049)	(0.056)	(0.489)	(0.265)	(0.239)
Age	-0.001	-0.001	-0.014	-0.009	
	(0.001)	(0.001)	(0.010)	(0.005)	
Type.of.Travel	-0.138***	-0.138***	-1.226***	-0.676***	-0.683***
	(0.045)	(0.053)	(0.431)	(0.237)	(0.184)
Class	0.017	0.017	0.111	0.076	
	(0.021)	(0.024)	(0.189)	(0.105)	
Flight.Distance	-0.00003*	-0.00003*	-0.0003	-0.0001	

	(0.00002)	(0.00002)	(0.0002)	(0.0001)	
Seat.comfort	-0.179***	-0.179***	-1.794***	-0.979***	-0.942***
	(0.025)	(0.028)	(0.269)	(0.146)	(0.139)
Food.and.drink	-0.081***	-0.081***	-0.917***	-0.488***	-0.458***
	(0.031)	(0.030)	(0.327)	(0.180)	(0.171)
Departure.Arrival.time.convenient	-0.115***	-0.115***	-1.085***	-0.636***	-0.616***
	(0.027)	(0.025)	(0.279)	(0.153)	(0.146)
Gate.location	-0.008	-0.008	-0.088	-0.116	-0.128
	(0.031)	(0.029)	(0.345)	(0.186)	(0.179)
Inflight.wifi.service	-0.015	-0.015	-0.042	0.001	
	(0.015)	(0.016)	(0.153)	(0.083)	
Inflight.entertainment	-0.003	-0.003	-0.022	0.032	0.012
	(0.031)	(0.032)	(0.329)	(0.178)	(0.172)
Online.support	0.025*	0.025	0.203	0.132	
	(0.015)	(0.015)	(0.147)	(0.081)	
Ease.of.Online.booking	0.071***	0.071***	0.518**	0.244**	0.343***
	(0.022)	(0.025)	(0.207)	(0.113)	(0.093)
On.board.service	-0.039	-0.039	-0.656*	-0.361*	-0.298
	(0.036)	(0.036)	(0.388)	(0.209)	(0.194)

Leg.room.service	0.027*	0.027**	0.241*	0.150*	
	(0.014)	(0.013)	(0.139)	(0.077)	
Baggage.handling	0.014	0.014	0.146	0.083	
	(0.017)	(0.017)	(0.167)	(0.092)	
Checkin.service	0.028**	0.028**	0.210*	0.128**	0.164***
	(0.012)	(0.012)	(0.117)	(0.065)	(0.062)
Cleanliness	-0.069**	-0.069**	-0.769**	-0.408**	-0.399**
	(0.032)	(0.033)	(0.338)	(0.183)	(0.174)
Online.boarding	-0.081**	-0.081**	-0.530	-0.282	-0.307*
	(0.035)	(0.040)	(0.347)	(0.190)	(0.177)
Arrival.Delay.in.Minutes	0.001	0.001	0.027	0.018	
	(0.004)	(0.004)	(0.037)	(0.021)	
Departure.Delay.in.Minutes	-0.004	-0.004	-0.055	-0.032	
	(0.004)	(0.004)	(0.038)	(0.021)	
Seat.comfort:Food.and.drink	0.080***	0.080***	0.772***	0.411***	0.401***
	(0.007)	(0.007)	(0.090)	(0.047)	(0.045)
Departure.Arrival.time.convenient:Gate.location	0.029***	0.029***	0.274***	0.165***	0.162***
	(0.009)	(0.008)	(0.092)	(0.050)	(0.047)
Food.and.drink:Gate.location	-0.043***	-0.043***	-0.384***	-0.201***	-0.199***

	(0.010)	(0.009)	(0.106)	(0.057)	(0.055)
On.board.service:Cleanliness	0.018*	0.018*	0.227**	0.121**	0.120**
	(0.010)	(0.010)	(0.102)	(0.055)	(0.051)
Inflight.entertainment:Online.boarding	0.028***	0.028***	0.220**	0.115**	0.123**
	(0.009)	(0.009)	(0.092)	(0.050)	(0.048)
Constant	0.538***	0.538***	1.899	0.981	1.191
	(0.184)	(0.190)	(1.800)	(0.994)	(0.865)
Observations	614		614	614	614
R^2	0.567				
Adjusted R^2	0.547				
Log Likelihood			-186.786	-187.903	-194.287
Akaike Inf. Crit.			429.572	431.807	426.575
Residual Std. Error	0.335 (df = 586)				
F Statistic	28.410*** (df = 27; 586)				
Note:					*p<0.1; **p<0.05; ***p<0.01

The figure above presents the initial OLS, logit and probit models, as well as the final model reached throughout this paper. As can be seen, there are still remaining variables that are statistically insignificant, i.e. Gate.location, Inflight.entertainment, On.board.service and Online.boarding. These variables are however part of other interactions, which proved to be statistically significant, therefore we decided based on likelihood ratio tests, not to remove them. When analyzing the probit model we cannot interpret the coefficients, we can only look whether the value is positive or negative, indicating the positive or negative influence on the target variable, without any information about the strength. The final model yields an Akaike information criterion of 426.575, being the smallest among other evaluated models, which also is an indication that the transformations done to the model variables improved the accuracy. While not printed out on the above figure, the R^2 statistic for the final model is equal to 0.4594722, which means that around 46% of the data variance is explained by the model.

Figure 11. Marginal effects of the variables in final probit.

Variable	**Marginal Effect**
Gender	-0.1187
Customer.type	0.2151
Type.of.travel	-0.1216
Seat.comfort	0.03646
Food.and.drink	0.02617
Departure.Arrival.time.convenient	-0.02727
Gate.location	-0.03635
Inflight.entertainment	0.07367
Ease.of.Online.booking	0.06101
On.board.service	0.02096
Checkin.service	0.02913
Cleanliness	-0.0006631
Online.boarding	0.01793

Before analyzing the marginal effects displayed on the Figure 11, let it be known that the interactions aren't variables per se and don't have separate effects. However, they are included in the effects of the specific variables that the interaction was made of. From the brief glance we can notice that two out of three remaining binary variables have negative effects, meaning that both Males and passengers traveling for business tend to be satisfied both around 12% less, while being a Loyal Customer increases the probability of being satisfied by 21.51%. Regarding the variables taken from a satisfaction survey the effects differ from one to another, which is surprising considering the assumption that the greater the score the greater the overall satisfaction should be. The probabilities are however small for the majority of variables, with the greatest effect having inflight.entertainment on the level of 7.367%. However the conducted analysis was enough to verify the hypothesis.

H1: Distance of travel has no influence on customer satisfaction.

Flight.distance was one of the variables that were removed from the model in the General to specific process. The variable proved to be statistically insignificant for each and every type of model assumed from the beginning of the analysis, therefore confirming the hypothesis that Distance of travel has no influence on customer satisfaction.

H2: Loyal customers are more forgiving and will be satisfied more often than the disloyal customers.

As can be seen in the model, customer loyalty is a statistically significant variable. Based on the marginal effects we can conclude that a change of disloyal to loyal customer increases the probability of the satisfaction by 21.51%. This supports our intuition and allows to confirm the hypothesis that loyal customers are more forgiving and will be satisfied more often than disloyal customers.

H3: Meeting customers' needs and requirements, and exceeding their expectations in quality of provided services lead to a higher probability of satisfaction.

Interestingly, some of the services, like Inflight.wifi.services were found to be statistically insignificant. This however does not mean that all services have no impact on satisfaction, as Seat.comfort, or Food.and.drink and others were statistically significant. It can also be observed that for all the services within the model, the marginal effects are positively influencing the satisfaction of customers. For example, an increase of rating for Seat.comfort by 1 increases

the satisfaction probability by 3.646%. This is satisfactory and allows for confirmation of the hypothesis that meeting customers' needs and requirements yields a higher probability of satisfaction.

Conclusion and future remarks

In our paper we conducted an econometric model on a satisfaction survey database to identify key influencing factors on the satisfaction from the travel. We started by cleaning the dataset, preparing the variables and creating interactions we believed were right. Then we compared the models using information criteria and R-squared and decided to proceed further with probit. After removing insignificant variables and conducting diagnostic tests we verified our hypothesis based on the marginal effects of the variables. All of our hypotheses were satisfied, which can be very telling about the ways in which the aviation industry can improve, as it is being more and more focused on the service and creating a pleasant experience.

Furthermore, with our preliminary analysis done, we see that there is potential for a lot of further analysis within the topic. For example we would like to see a more detailed database, where more information about flight duration was collected. It could be the case that for some carriers the duration of flight and satisfaction could have a meaningful relation, however it is blurred out by the amount of other data collected. It would also be interesting to be able to map the average ticket prices, and how full and frequent the flight is. This could show underlying effects that are not detected with the dataset at our hand.

Bibliography

Chen, G., & Tsurumi, H. (2010). *Probit and Logit Model Selection. Communications in Statistics - Theory and Methods,* 40(1), p. 159-175.

Hayadi, B. H., Kim, J. M., Hulliyah, K., & Sukmana, H. T. (2021). *Predicting Airline Passenger Satisfaction with Classification Algorithms. International Journal of Informatics and Information System,* 4(1), p. 82-94.

Hussain, R., Al Nasser, A., & Hussain, Y. K. (2015). *Service quality and customer satisfaction of a UAE-based airline: An empirical investigation. Journal of Air Transport Management,* 42, p. 167-175.

Jiang, H., & Zhang, Y. (2016). *An investigation of service quality, customer satisfaction and loyalty in China's airline market. Journal of Air Transport Management,* 57, p. 80-88.

Namukasa, J. (2013). *The influence of airline service quality on passenger satisfaction and loyalty. The TQM Journal*, 25(5), p. 520-532.

Appendix

```
#########################################################################
###   Advanced Econometrics - Airline Customer satisfaction
###
###      Aleksander Wieliński  -     420 272
###      Jakub Gazda           -     419 272
###
#########################################################################

### LIBRARIES
#########################################################################

#install.packages("dplyr")
library("dplyr")
#install.packages("corrplot")
library('corrplot')
#install.packages("margins")
library("margins")
#install.packages("plm")
library("plm")
#install.packages("ResourceSelection")
library("ResourceSelection")
#install.packages("ggplot2")
library("ggplot2")
#install.packages("lattice")
library("lattice")
#install.packages("caret")
library("caret")
#install.packages("lmtest")
library("lmtest")
#install.packages("DescTools")
#install.packages("RDCOMClient", repos="http://www.omegahat.net/R")
library("DescTools")
#install.packages('aods3')
library("aods3")

#########################################################################


### DATA LOADING AND PREPARATION

setwd(getwd())
data <- as.data.frame(read.csv("Invistico_Airline.csv"))

# Random seed - this ensures consistent data across devices
set.seed(123)
```

```
# Splitting the dataset while maintaining the target variable distribution.
# This is done to save memory and computational resources and allow for a fast analysis.
trainIndex <- createDataPartition(data$satisfaction, p = 0.005, list = FALSE, times = 1)

# Create the smaller subset by selecting the instances based on the indices
data <- data[trainIndex, ]

# Missing values removal
na_count <- sum(is.na(data))
print(na_count)
data<-na.omit(data)
#View(data)

# Binary encoding
data$satisfaction<-ifelse(data$satisfaction == "satisfied", 1,0)
data$Gender<-ifelse(data$Gender == "Male", 1,0)
data$Customer.Type<-ifelse(data$Customer.Type == "Loyal Customer", 1,0)
data$Type.of.Travel<-ifelse(data$Type.of.Travel == "Personal Travel", 1,0)
data$Class <- as.integer(factor(data$Class, levels = c("Eco", "Eco Plus", "Business"), labels = c(1, 2, 3)))

variables <- c("Seat.comfort", "Departure.Arrival.time.convenient", "Food.and.drink",
               "Gate.location", "Inflight.wifi.service", "Inflight.entertainment",
               "Online.support", "Ease.of.Online.booking","On.board.service", "Leg.room.service",
               "Baggage.handling", "Checkin.service", "Cleanliness", "Online.boarding")

str(data)
summary(data)

x <- data[(data$Age==0), ]
x

### Exploratory Data Analysis

hist(data$Departure.Delay.in.Minutes)

hist(log1p(data$Departure.Delay.in.Minutes))

hist(BoxCox(data$Departure.Delay.in.Minutes, 0))


hist(data$Arrival.Delay.in.Minutes)

hist(log1p(data$Arrival.Delay.in.Minutes))

hist(BoxCox(data$Arrival.Delay.in.Minutes, 0))

# Heatmap - correlations between variables
par(mfrow = c(1,1))
cor_matrix <- cor(data)
corrplot(cor_matrix, method = "color", tl.col = "black", tl.srt = 0)

#print_correlation_table <- function(data, variables) {
#  cor_matrix <- cor(data[, variables])
#  print(cor_matrix)
#}

par(mar = c(2, 2, 2, 2))
columns1 <- c("Age", "Flight.Distance", "Departure.Delay.in.Minutes", "Arrival.Delay.in.Minutes")
```

```
par(mfrow=c(2,2))

# Histograms for the columns in columns1
for (col in columns1) {
  hist(data[[col]], main = col, xlab = "", col = "lightblue")
}
# Boxplots for the columns in columns1
for (col in columns1) {
  boxplot(data[[col]], main = col, col = "lightblue")
}

columns2 <- c("satisfaction", "Gender", "Customer.Type","Type.of.Travel", "Class",
              "Seat.comfort", "Departure.Arrival.time.convenient","Food.and.drink",
              "Gate.location","Inflight.wifi.service", "Inflight.entertainment", "Online.support",
              "Ease.of.Online.booking","On.board.service", "Leg.room.service", "Baggage.handling",
              "Checkin.service", "Cleanliness","Online.boarding")

par(mfrow = c(4, 5))

# Bar plots for the columns in columns2
for (col in columns2) {
  barplot(table(data[[col]]), main = col, xlab = "", col = "lightblue")
}

for (col in c("Flight.Distance")) { #, "Departure.Delay.in.Minutes", "Arrival.Delay.in.Minutes"
  threshold <- quantile(data[[col]], 0.95)
  data <- data[data[[col]] <= threshold, ]
}

columns3 <- c("Age", "Flight.Distance", "Departure.Delay.in.Minutes", "Arrival.Delay.in.Minutes")
par(mfrow=c(2,2))
# Histograms for the columns in columns1
for (col in columns3) {
  hist(data[[col]], main = col, xlab = "", col = "lightblue")
}

par(mfrow = c(1,1))
# A non linear relation is introduced in order to normalize the distribution
# of those 2 variables.

data$Departure.Delay.in.Minutes <- log(data$Departure.Delay.in.Minutes+0.001)
data$Arrival.Delay.in.Minutes <- log(data$Arrival.Delay.in.Minutes+0.001)

### Assesing models

source("linktest.R")

## OLS

lpm <- lm(satisfaction~Gender+Customer.Type+Age+Type.of.Travel+Class+Flight.Distance
          +Seat.comfort*Food.and.drink+Departure.Arrival.time.convenient*Gate.location
          +Food.and.drink*Gate.location+Inflight.wifi.service+Inflight.entertainment
          +Online.support+Ease.of.Online.booking+On.board.service+Leg.room.service
          +Baggage.handling+Checkin.service+Cleanliness*On.board.service
          +Online.boarding*Inflight.entertainment+Arrival.Delay.in.Minutes
          +Departure.Delay.in.Minutes, data=data)

summary(lpm)
```

```
PseudoR2(lpm, "all")

# specification test
resettest(lpm, power=2:3, type="fitted")

# heteroscedasticity
lpm.residuals = lpm$residuals
plot(lpm.residuals~Gender, data=data)
plot(lpm.residuals~Customer.Type, data=data)

plot(lpm.residuals~Gender+Customer.Type+Age+Type.of.Travel+Class+Flight.Distance
    +Seat.comfort*Food.and.drink+Departure.Arrival.time.convenient*Gate.location
    +Food.and.drink*Gate.location+Inflight.wifi.service+Inflight.entertainment
    +Online.support+Ease.of.Online.booking+On.board.service+Leg.room.service
    +Baggage.handling+Checkin.service+Cleanliness*On.board.service
    +Online.boarding*Inflight.entertainment+Arrival.Delay.in.Minutes
    +Departure.Delay.in.Minutes, data=data)

bptest(lpm.residuals~Gender, data=data)
bptest(lpm.residuals~Gender+Customer.Type, data=data)

bptest(lpm.residuals~Gender+Customer.Type+Age+Type.of.Travel+Class+Flight.Distance
      +Seat.comfort*Food.and.drink+Departure.Arrival.time.convenient*Gate.location
      +Food.and.drink*Gate.location+Inflight.wifi.service+Inflight.entertainment
      +Online.support+Ease.of.Online.booking+On.board.service+Leg.room.service
      +Baggage.handling+Checkin.service+Cleanliness*On.board.service
      +Online.boarding*Inflight.entertainment+Arrival.Delay.in.Minutes
      +Departure.Delay.in.Minutes, data=data)



# White's estimator of the variance-covariane matrix
robust_vcov = vcovHC(lpm, data = olympics, type = "HC")
coeftest(lpm, vcov.=robust_vcov)

# to compare the simple lpm and the one with a robust vcov matrix
#install.packages("stargazer")
library("stargazer")
robust.lpm = coeftest(lpm, vcov.=robust_vcov)
#stargazer(lpm, robust.lpm, type="html",
#          out="C:/Users/Aleksander.Wielinski/Desktop/Documents/0.Pers-UNI/__AE/A_Econometrics-main/A_Econometrics-main/lpm-stargaze.html")

## Logit Modelling

mylogit <- glm(satisfaction~Gender+Customer.Type+Age+Type.of.Travel+Class+Flight.Distance
               +Seat.comfort*Food.and.drink+Departure.Arrival.time.convenient*Gate.location
               +Food.and.drink*Gate.location+Inflight.wifi.service+Inflight.entertainment
               +Online.support+Ease.of.Online.booking+On.board.service+Leg.room.service
               +Baggage.handling+Checkin.service+Cleanliness*On.board.service
               +Online.boarding*Inflight.entertainment+Arrival.Delay.in.Minutes
               +Departure.Delay.in.Minutes, data=data,
               family=binomial(link="logit"))

# Model Summary
summary(mylogit)
logit_summary <- summary(mylogit)

# R^2 statistics
r_squared_logit <- logit_summary$deviance/logit_summary$null.deviance
```

```
print(r_squared_logit)
PseudoR2(mylogit, "all")

# Linktest - yhat significant | yhat2 insignificant -> cannot reject H0
linktest_result_logit = linktest(mylogit)
summary(linktest_result_logit)

# Stargaze

stargazer(mylogit, type="text")

## Probit Modelling

#myprobit <-
glm(satisfaction~Gender+Customer.Type+Age+Type.of.Travel+Class+Flight.Distance+Seat.comfort+Departure.Arrival.time.convenient+Food.and.drink+Gate.
location+Inflight.wifi.service+Inflight.entertainment+Online.support+Ease.of.Online.booking+On.board.service+Leg.room.service+Baggage.handling+Checkin
.service+Online.boarding+Arrival.Delay.in.Minutes+Departure.Delay.in.Minutes, data=data,
#                family=binomial(link="probit"))
#no diff when 1,0 are 50-50, information criteria: we should use AIC to be sure, BIC(SBC) - its better, the lower the value the better
myprobit <- glm(satisfaction~Gender+Customer.Type+Age+Type.of.Travel+Class
                +Flight.Distance+Seat.comfort*Food.and.drink
                +Departure.Arrival.time.convenient*Gate.location+Food.and.drink*Gate.location
                +Inflight.wifi.service+Inflight.entertainment+Online.support
                +Ease.of.Online.booking+On.board.service+Leg.room.service
                +Baggage.handling+Checkin.service+Cleanliness*On.board.service
                +Online.boarding*Inflight.entertainment+Arrival.Delay.in.Minutes
                +Departure.Delay.in.Minutes, data=data,
                family=binomial(link="probit"))

# Model Summary
summary(myprobit)
probit_summary <- summary(myprobit)

# R^2 statistics
r_squared_probit <- probit_summary$deviance/probit_summary$null.deviance
print(r_squared_probit)
PseudoR2(myprobit, "all")
#can interprate McKelvy.Zavoina if the latent variable is observed the model explainx x% of observations, count is the correct R2 explaining x% of the
observation, adj.count is saying % ofcorrectly predicted observations with given variance

# Linktest - yhat significant | yhat2 insignificant -> cannot reject H0
linktest_result_probit = linktest(myprobit)

summary(linktest_result_probit)

stargazer(myprobit, type="html", out="C:/Users/48796/OneDrive/Pulpit/STUDIA/ROK 4/SEM 2/A. Econometrics/project/myprobit1.html")

### Diagnostics

# can also use Wald test for beta2=beta3=0
#H <- rbind(c(0,1,0,0), c(0,0,1,0))
# h %*% coef(dem.probit)

# DECLARE TERMS
#?wald.test
#wald_results = wald.test(myprobit, )
#wald_results
```

```
# Goodness of fit test
gof_results = gof(myprobit)
gof_results


### General to Specific procedure

# for 0,005 of dataset
# general to specific thatway= h0 is beta=0(for additional variables) and with p-value<0.05 we reject h0. Joint insignificance of all variables test against null
null_probit = glm(satisfaction~1, data=data, family=binomial(link="probit"))
lrtest(myprobit, null_probit)
#bez inflight wifi service
myprobit1 <-
glm(satisfaction~Gender+Customer.Type+Age+Type.of.Travel+Class+Flight.Distance+Seat.comfort*Food.and.drink+Departure.Arrival.time.convenient*Gate.
location+Food.and.drink*Gate.location+Inflight.entertainment+Online.support+Ease.of.Online.booking+On.board.service+Leg.room.service+Baggage.handling
+Checkin.service+Cleanliness*On.board.service+Online.boarding*Inflight.entertainment+Arrival.Delay.in.Minutes+Departure.Delay.in.Minutes, data=data,
        family=binomial(link="probit"))
lrtest(myprobit, myprobit1)
summary(myprobit1)
#bez inflight.entertainment
myprobit2 <-
glm(satisfaction~Gender+Customer.Type+Age+Type.of.Travel+Class+Flight.Distance+Seat.comfort*Food.and.drink+Departure.Arrival.time.convenient*Gate.
location+Food.and.drink*Gate.location+Online.support+Ease.of.Online.booking+On.board.service+Leg.room.service+Baggage.handling+Checkin.service+Clea
nliness*On.board.service+Online.boarding+Arrival.Delay.in.Minutes+Departure.Delay.in.Minutes, data=data,
        family=binomial(link="probit"))
lrtest(myprobit1, myprobit2)
### the change was too significant, go to the next biggest pvalue

#bez gate location
myprobit3 <-
glm(satisfaction~Gender+Customer.Type+Age+Type.of.Travel+Class+Flight.Distance+Seat.comfort*Food.and.drink+Departure.Arrival.time.convenient+Food.
and.drink+Inflight.entertainment+Online.support+Ease.of.Online.booking+On.board.service+Leg.room.service+Baggage.handling+Checkin.service+Cleanlines
s*On.board.service+Online.boarding*Inflight.entertainment+Arrival.Delay.in.Minutes+Departure.Delay.in.Minutes, data=data,
        family=binomial(link="probit"))
lrtest(myprobit1, myprobit3)
### the change was too significant, go to the next biggest pvalue

#bez class
myprobit4 <-
glm(satisfaction~Gender+Customer.Type+Age+Type.of.Travel+Flight.Distance+Seat.comfort*Food.and.drink+Departure.Arrival.time.convenient*Gate.locatio
n+Food.and.drink*Gate.location+Inflight.entertainment+Online.support+Ease.of.Online.booking+On.board.service+Leg.room.service+Baggage.handling+Chec
kin.service+Cleanliness*On.board.service+Online.boarding*Inflight.entertainment+Arrival.Delay.in.Minutes+Departure.Delay.in.Minutes, data=data,
        family=binomial(link="probit"))
lrtest(myprobit1, myprobit4)
summary(myprobit4)
# bez arrival delay
myprobit5 <-
glm(satisfaction~Gender+Customer.Type+Age+Type.of.Travel+Flight.Distance+Seat.comfort*Food.and.drink+Departure.Arrival.time.convenient*Gate.locatio
n+Food.and.drink*Gate.location+Inflight.entertainment+Online.support+Ease.of.Online.booking+On.board.service+Leg.room.service+Baggage.handling+Chec
kin.service+Cleanliness*On.board.service+Online.boarding*Inflight.entertainment+Departure.Delay.in.Minutes, data=data,
        family=binomial(link="probit"))
lrtest(myprobit4, myprobit5)
summary(myprobit5)
# bez baggage handling
myprobit6 <-
glm(satisfaction~Gender+Customer.Type+Age+Type.of.Travel+Flight.Distance+Seat.comfort*Food.and.drink+Departure.Arrival.time.convenient*Gate.locatio
n+Food.and.drink*Gate.location+Inflight.entertainment+Online.support+Ease.of.Online.booking+On.board.service+Leg.room.service+Checkin.service+Cleanli
ness*On.board.service+Online.boarding*Inflight.entertainment+Departure.Delay.in.Minutes, data=data,
```

```
        family=binomial(link="probit"))
lrtest(myprobit5, myprobit6)
summary(myprobit6)
# bez departure delay in min
myprobit7 <-
glm(satisfaction~Gender+Customer.Type+Age+Type.of.Travel+Flight.Distance+Seat.comfort*Food.and.drink+Departure.Arrival.time.convenient*Gate.location+Food.and.drink*Gate.location+Inflight.entertainment+Online.support+Ease.of.Online.booking+On.board.service+Leg.room.service+Checkin.service+Cleanliness*On.board.service+Online.boarding*Inflight.entertainment, data=data,
        family=binomial(link="probit"))
lrtest(myprobit6, myprobit7)
summary(myprobit7)
#bez online support
myprobit8 <-
glm(satisfaction~Gender+Customer.Type+Age+Type.of.Travel+Flight.Distance+Seat.comfort*Food.and.drink+Departure.Arrival.time.convenient*Gate.location+Food.and.drink*Gate.location+Inflight.entertainment+Ease.of.Online.booking+On.board.service+Leg.room.service+Checkin.service+Cleanliness*On.board.service+Online.boarding*Inflight.entertainment, data=data,
        family=binomial(link="probit"))
lrtest(myprobit7, myprobit8)
summary(myprobit8)
#bez online boarding
myprobit9 <-
glm(satisfaction~Gender+Customer.Type+Age+Type.of.Travel+Flight.Distance+Seat.comfort*Food.and.drink+Departure.Arrival.time.convenient*Gate.location+Food.and.drink*Gate.location+Inflight.entertainment+Ease.of.Online.booking+On.board.service+Leg.room.service+Checkin.service+Cleanliness*On.board.service, data=data,
        family=binomial(link="probit"))
lrtest(myprobit8, myprobit9)
### the change was too significant, go to the next biggest pvalue

#bez age
myprobit10 <-
glm(satisfaction~Gender+Customer.Type+Type.of.Travel+Flight.Distance+Seat.comfort*Food.and.drink+Departure.Arrival.time.convenient*Gate.location+Food.and.drink*Gate.location+Inflight.entertainment+Ease.of.Online.booking+On.board.service+Leg.room.service+Checkin.service+Cleanliness*On.board.service+Online.boarding*Inflight.entertainment, data=data,
        family=binomial(link="probit"))
lrtest(myprobit8, myprobit10)
summary(myprobit10)
#bez flight distance
myprobit11 <-
glm(satisfaction~Gender+Customer.Type+Type.of.Travel+Seat.comfort*Food.and.drink+Departure.Arrival.time.convenient*Gate.location+Food.and.drink*Gate.location+Inflight.entertainment+Ease.of.Online.booking+On.board.service+Leg.room.service+Checkin.service+Cleanliness*On.board.service+Online.boarding*Inflight.entertainment, data=data,
        family=binomial(link="probit"))
lrtest(myprobit10, myprobit11)
summary(myprobit11)
# bez on board service
myprobit12 <-
glm(satisfaction~Gender+Customer.Type+Type.of.Travel+Seat.comfort*Food.and.drink+Departure.Arrival.time.convenient*Gate.location+Food.and.drink*Gate.location+Inflight.entertainment+Ease.of.Online.booking+On.board.service+Leg.room.service+Checkin.service+Cleanliness*On.board.service+Online.boarding*Inflight.entertainment, data=data,
        family=binomial(link="probit"))
lrtest(myprobit11, myprobit12)
### the change was too significant, go to the next biggest pvalue

#bez leg room service
myprobit13 <-
glm(satisfaction~Gender+Customer.Type+Type.of.Travel+Seat.comfort*Food.and.drink+Departure.Arrival.time.convenient*Gate.location+Food.and.drink*Gate.location+Inflight.entertainment+Ease.of.Online.booking+On.board.service+Checkin.service+Cleanliness*On.board.service+Online.boarding*Inflight.entertainment, data=data,
```

```
            family=binomial(link="probit"))
lrtest(myprobit11, myprobit13)
summary(myprobit13)

# Link test
linktest_result = linktest(myprobit13)
stargazer(linktest_result, type="text", out="C:/Users/48796/OneDrive/Pulpit/STUDIA/ROK 4/SEM 2/A. Econometrics/project/linktest_13.html")

# Model information
model_summary13 <- summary(myprobit13)
r_squared13 <- model_summary13$deviance/model_summary13$null.deviance
print(r_squared13)
PseudoR2(myprobit13, "all")

# Breusch-Pagan test
bptest(myprobit13, data=data)

# Goodnes of fit tests
gof.results = gof(myprobit13)
gof.results

predicted_probs <- predict(myprobit13, type = "response")

hosmer_lemeshow <- hoslem.test(data$satisfaction, predicted_probs)
stargazer(hosmer_lemeshow, type="text", out="C:/Users/48796/OneDrive/Pulpit/STUDIA/ROK 4/SEM 2/A. Econometrics/project/HL_13.html")

print(hosmer_lemeshow)
stargazer(myprobit, myprobit13, type="html",
      out="C:/Users/48796/OneDrive/Pulpit/STUDIA/ROK 4/SEM 2/A. Econometrics/project/probit1vs13.html")
stargazer(lpm, mylogit, myprobit, myprobit13, type="text")

stargazer(lpm, robust.lpm, mylogit, myprobit, myprobit13, type="html",
      out="C:/Users/48796/OneDrive/Pulpit/STUDIA/ROK 4/SEM 2/A. Econometrics/project/allmodels.html")

info_criterion <- data.frame(model = c("robust.lpm", "mylogit", "myprobit"),
                     AIC = c(AIC(robust.lpm), AIC(mylogit), AIC(myprobit)),
                     BIC = c(BIC(robust.lpm), BIC(mylogit), BIC(myprobit))
)

info_criterion

info_criterion <- data.frame(model = c("robust.lpm", "mylogit", "myprobit", "myprobit13"),
                     AIC = c(AIC(robust.lpm), AIC(mylogit), AIC(myprobit), AIC(myprobit13)),
                     BIC = c(BIC(robust.lpm), BIC(mylogit), BIC(myprobit), BIC(myprobit13))
)

info_criterion

# Calculate marginal effects
marginal_effects <- margins(myprobit13, data = data)
# Print the marginal effects
print(marginal_effects)
stargazer(marginal_effects, type="text")
stargazer(summary(marginal_effects), type="text")
```