Factors Affecting the Conversion Rate in the Flight Comparison Industry: A Logistic Regression Approach

YOUSEF AHO

JOHANNES PERSSON
Factors Affecting the Conversion Rate in the Flight Comparison Industry: A Logistic Regression Approach

YOUSEF AHO
JOHANNES PERSSON
Abstract

Using logistic regression, we aim to construct a model to examine the factors that are most influential in affecting user behavior on the flight comparison site flygresor.se. The factors examined were number of adults, number of children, number of stops on the inbound trip, number of stops on the outbound trip, number of days between the search date and the departure date and number of search results displayed for the user. The data sample, collected during a one-week period, was taken from Flygresor and consisted of trips to or from Sweden, made within Europe, excluding Nordic countries, and made more than six days before departure. To find the variables which best explain the user behavior, variable selection methods were used along with hypothesis testing. Also, multicollinearity analysis and residual analysis were performed to evaluate the final model. The result showed that the factor number of children had no significant impact on the conversion rate, while the remaining factors had a high impact. The final model has a high predictive ability on the user’s propensity to select a certain flight.
Sammanfattning

Med hjälp av logistisk regression ämnar vi att ta fram en modell som beskriver vilka faktorer som påverkar användarbeteendet på prisjämförelsesajten flygresor.se. Faktorerna som analyserades var antalet vuxna, antalet barn, antalet mellanlandningar på utgående resa, antalet mellanlandningar på ingående resa, antalet dagar mellan sökdatum och avresedatum och antalet sökresultat som visas för användaren. Datat som användes var taget från Flygresor under en veckas tid, och bestod av resor till och från Sverige, gjorda inom Europa men inte till nordiska länder, med mer än sex dagar till avresedatum. För att hitta de variabler som förklarar användarbeteendet mest, användes olika urvalsmetoder och hypotesprövningar samt olika evalueringsmetoder som residualanalys. Resultatet visade att variabeln antalet barn inte hade någon signifikant påverkan på konverteringsgraden, medan resterande faktorer hade en hög påverkan. Den slutgiltiga modellen har en hög förmåga att förutse användarens sannolikhet att välja en specifik flygresa.
Acknowledgements

We would like to thank our supervisor at the department of mathematical statistics at KTH, Per Jörgen Säve-Söderbergh, for his great support and valuable advice. In addition, we would also like to give a special thanks to Mattias Nyman and Robin Dahlberg at Flygresor and their associates for providing us with data, relevant information and support.
## Contents

1 Introduction .............................. 1
   1.1 Background ................................. 1
   1.2 Motivation and Impact ................. 2
   1.3 Research Question ...................... 2
   1.4 Scope and Limitations ................. 3
   1.5 Earlier Research ...................... 4
   1.6 Feasibility ......................... 4

2 Theoretical Framework ............... 5
   2.1 Regression Analysis ......... 5
   2.2 Logistic Regression Analysis ...... 5
      2.2.1 Bernoulli Distribution ..... 5
      2.2.2 Logit Transformation ...... 6
      2.2.3 Logistic Regression Model ... 6
      2.2.4 Maximum Likelihood ....... 7
   2.3 Model Evaluation and Diagnostics ... 7
      2.3.1 Multicollinearity ............ 7
      2.3.2 Akaike Information Criterion ... 8
      2.3.3 Bayesian Information Criterion ... 8
      2.3.4 Hypothesis Testing ...... 8
      2.3.5 Wald Inference .......... 9
      2.3.6 Confidence Interval Estimation ... 9
      2.3.7 Cook's Distance .... 9
      2.3.8 ROC Curve .......... 10
   2.4 Variable Selection .......... 10
      2.4.1 Stepwise Regression .... 10
      2.4.2 Lasso Regression .... 10
      2.4.3 k-fold Cross-Validation .... 11

3 Methodology ...................... 12
   3.1 Description of Data ......... 12
      3.1.1 Data Collection ........ 12
      3.1.2 Data Management ........ 12
      3.1.3 Response Variable ........ 12
      3.1.4 Regressor Variables .... 13
   3.2 Model Creation ............. 13
      3.2.1 Multicollinearity .... 13
      3.2.2 Variable Selection .... 14
   3.3 Model Evaluation .......... 14
      3.3.1 Residual Analysis .... 14
      3.3.2 Wald Inference .... 14
      3.3.3 Confidence Interval Estimation ... 14
      3.3.4 ROC Curve ........ 14
4 Results

4.1 Data Selection ............................................ 15
4.2 Full Model .................................................. 15
  4.2.1 Multicollinearity ...................................... 15
  4.2.2 Confidence Interval Estimation ......................... 16
  4.2.3 Cook’s Distance ........................................ 17
4.3 Variable Selection ........................................ 17
  4.3.1 Forward Selection and Backward Elimination .......... 17
  4.3.2 Lasso Regression with k-fold Cross Validation ....... 18
  4.3.3 Variable Influence ..................................... 18
4.4 Reduced Model ............................................ 19
  4.4.1 Multicollinearity ...................................... 19
  4.4.2 Confidence Interval Estimation ......................... 20
  4.4.3 Cook’s Distance ........................................ 20
4.5 Model Evaluation .......................................... 21
  4.5.1 ROC Curve and AUC ..................................... 21

5 Discussion .................................................. 22

5.1 Model Creation ............................................ 22
  5.1.1 Choice of Regression Model ............................. 22
  5.1.2 Candidate Variables .................................. 22
  5.1.3 Reduced Model ......................................... 23
5.2 Model Accuracy ............................................ 23
  5.2.1 Cook’s Distance ........................................ 23
  5.2.2 Predictive Ability ..................................... 23
5.3 Model Interpretation ....................................... 24
  5.3.1 Variables and Coefficients ............................. 24
5.4 Model Implication ......................................... 25
5.5 Model Application ......................................... 26
  5.5.1 Usage Guidelines ...................................... 26
5.6 Further development ...................................... 26

6 Conclusions ................................................. 27
1. Introduction

This project has been performed with the flight comparison site flygresor.se. The company compares the prices on flight tickets from 34 travel agencies and more than 900 airlines.

1.1. Background

Today, we are in the midst of a digital revolution. In the last few years, many companies have turned to being primarily data driven in their operations [12]. For companies to adapt to this new way of doing business, they all have the same goal: to collect as much data as possible. For several years, this has been the primary focus [12]. Now, however, the focus is to actually derive some useful information from this data. A study from the International Data Corporation shows that only 1% of all data that has been collected is used for analysis [6]. Since much of this data is ready to be analyzed, the focus should be on deriving information from the data [2]. There are two reasons for this focus. Firstly, data collection can be resource intensive, and is inefficient if the data is not analyzed. Secondly, analyzing data gives direction for future data collection. There is a high risk that the wrong data is being collected if knowledge of the data is lacking. Therefore, projects aiming solely to extract information from existing data are highly relevant and important.

Another important development that we see in the business sphere today is a shift from a producer-driven market to a consumer-driven market [11]. Companies used to be able to get by with decent customer support and maintaining their customers simply by doing what they had always done. Now, the customer is in control. A higher availability of information and larger freedom to choose has brought a new level of competitiveness. Because of increased competition, the customer can make more demands. Consequently, the producers benefit from having as much information about the customer’s needs as possible. The company that can find the most information about its customers can more easily satisfy their requirements, and control the market. Using customer data to shape the offerings allows the producer to tailor the offer to each individual customer, or to a specific group of customers. This is especially important for internet-based companies.

The flight industry has over the past decades gone through a transformation since the deregulation of the industry in 1970. The supply and demand on air travel has increased drastically over the years because of this [3]. Today, we encounter different kinds of actors in the market. For instance, there are a lot more low-budget airlines in the market, which has decreased the price of air travel [16]. Since the competition in the market has increased, airlines utilize travel agencies to promote their company. These travel agencies uses flight comparison sites, like Flygresor, to promote themselves. Since the supply has increased, consumers rely a lot more on flight comparison sites to increase their
knowledge about the market (see figure 1) [15].

1.2. Motivation and Impact
In order for Flygresor to increase their service in the market and hence their competitiveness, they need to increase their understanding of the customers. Flygresor has a large amount of data that has not been analyzed. To make models both for converting the current Big Data to Smart Data and to construct prediction models is therefore of high value. The conclusions for this project will determine future strategy and could also lay the groundwork for other thesis projects. Specifically, Flygresor will be able to further improve the efficiency and quality of their search results. This will allow Flygresor to increase their market share. The project will primarily be of value to Flygresor, and secondarily for other researchers.

1.3. Research Question
The research question can be formulated as:

- Which factors affect the conversion rate at the flight comparison website flygresor.se the most?

The factors that will be examined are:

- Number of adults (adults)
- Number of children (children)
- Number of stops on the outbound trip (outboundStops)
- Number of stops on the inbound trip (inboundStops)
- Number of days between the search date and the departure date (daysAhead)
- Number of search results displayed for the user (flightAlternatives)
1.4. Scope and Limitations

To fit the scope of this project, we need to apply both qualitative and quantitative limitations. To make the data set fit for our research question, we want a homogeneous set of search results, given the information that is already known by Flygresor and their associates. Information given by Supersaver, an online travel agency (OTA) owned by Flygresor, revealed that there are two primary factors known to affect the conversion rate for searches: the time from booking a flight until the departure, and the region in which the flights are booked. These factors are also confirmed by Andrez Martinez et al. and Szopiński et al. [1, 18]. The information also revealed that business travellers behave differently in comparison with regular travellers. Specifically, the majority of Swedish business travellers only book Nordic flights within 0-6 days before departure.

For this study, we analyze flights to or from Sweden. In order to obtain a homogeneous data set we only include trips made within Europe but outside of the Nordic region (excluding Iceland). We also exclude trips booked less than seven days before departure. This separates the majority of business travellers from leisure travellers.

Within this segment we also decide to make further limitations. When booking a flight via Flygresor, there are three flight options: round trip, one-way, and open jaw (leaving from one destination and returning from another). One-way and open jaw trips are often booked in a different way than round trips. We therefore limit the scope to round trips. Since Flygresor is an international company, searches are made from the Swedish domain flygresor.se, as well as from other domains. Primarily, searches are made from the Swedish one. Since people in other countries book for different regions and might also behave differently, we choose to limit the scope to the Swedish domain. Finally, there are different types of tickets: Economy, Economy Plus, Business and First Class. According to Supersaver, customers who book Business and First Class tickets behave differently than those who book Economy and Economy Plus tickets. Since the lion’s share of the tickets are Economy, we choose to only analyze these bookings.
1.5. **Earlier Research**

Extensive research has been conducted on analyzing the purchase behavior, the factors that affect a purchase of a flight. The majority of the research made, can be divided into two main groups. The first is qualitative research made on customer behavior, with factors such as *perceived usefulness*, *ease of use* and *performance expectancy*. The second one is quantitative research focusing on actual conversion rate factors, such as *month of departure*, *days in advance* and *morning flight*.

One qualitative study shows that *habit*, *price saving*, *performance expectancy* and *facilitating conditions* are the most important factors, in that order [5]. Another one deems *perceived usefulness*, *ease of use* and *attitude towards online shopping* to be the most influential factors on user behavior [17]. We note that the qualitative method and the difficulty to measure many of these variables renders much of this groundwork unfit to build our research on. The only variable we could measure is *price saving*. However, even this is analyzed in a perceptive way, and does not provide us with much information.

Analyzing the quantitative studies, we notice that their findings provides us with valuable information. For instance, one study concluded that the propensity to buy a flight in advance decreases as the price increases. However, the propensity increases during vacation time, implying a lower price-sensitivity during these periods [4]. Another study shows a clear relationship between the *time of departure* and the price level of the flight, as well as a relationship between the *destination* and the price [1]. Since the price for a flight is largely based on the demand for it (using price discrimination), there is most likely a relationship between the conversion rate and these variables as well [5]. This was also confirmed by Supersaver.

1.6. **Feasibility**

Our main goal with this project is to find a conclusion from the given data that gives Flygresor a direction in their marketing and strategy. Rather than to go in to detail on every regressor variable, we want to get a general overview of the five or six most relevant and influential factors in customer decision making. We expect a high risk of delay in data handling, due to the large data sample size and our level of analyzing experience. Information regarding segmentation, pricing models and customer behavior, which is not given in the data but given by individuals at the company and its subsidiaries, is assumed to be accurate and relevant. In order to convert the given data to a form that is sufficient to work with, we need to gain knowledge in data management and the database the company is working with. This will be mainly done using available information online and support from Flygresor.
2. Theoretical Framework

In this section, we present all mathematical framework needed to complete this project.

2.1. Regression Analysis

Regression analysis is a very common technique for analyzing multi-factor data such as our data set. It utilizes the process of using an equation to express the relationship between a variable of interest (the response variable) and a set of related regressor variables \([14, p. XIII]\). Regression analysis also includes treatment of practical problems such as multicollinearity and remote data points, that arise when the mathematical models are applied on real data. This will be presented more thoroughly in the sections below.

2.2. Logistic Regression Analysis

Logistic regression is a technique used when the response variable is binary, i.e. it can only take two values, namely 0 and 1. Generally, 0 is called a failure and 1 is called a success. In logistic regression homoscedasticity is not required, the observations are independent and the error terms are uncorrelated with mean 0. \([14, p. 421-425]\)

2.2.1. Bernoulli Distribution

Given that we have a model in the following form:

\[ y_i = x_i^T \beta + \epsilon_i, \quad i = 1, \ldots, n \]  

where \( y_i \) is the response for the i:th observation, \( x_i^T = (1, x_{i1}, x_{i2}, \ldots, x_{ik}) \) are the regressor values for the i:th observation, \( \beta = (\beta_0, \beta_1, \ldots, \beta_k) \) are the coefficient values and \( \epsilon_i \) is the error component. We assume, in the logistic regression model, that the response variable \( y_i \) is a Bernoulli random variable with the following probability distribution:

\[ y_i = \begin{cases} 1, & \text{with probability } P(y_i = 1) = \pi_i \\ 0, & \text{with probability } P(y_i = 0) = 1 - \pi_i \end{cases} \]

Since \( y_i \) is Bernoulli distributed, the probability density function is given by:

\[ f_i(y_i) = \pi_i^{y_i} (1 - \pi_i)^{1-y_i}, \quad i = 1, \ldots, n \]

Knowing that the expected value of the error component for each observation \( i \) is 0, the expected value of \( y_i \) is:

\[ E(y_i) = 1 \cdot \pi_i + 0 \cdot (1 - \pi_i) = \pi_i = x_i^T \cdot \beta \]

This indicates that the expected response is the probability that the response variable takes on the value 1. \([14, p. 422-424]\)
2.2.2. Logit Transformation

In the multiple regression model we assume that the error component is normally distributed with mean 0 and constant variance $\sigma^2$. Because the response variable in the logistic regression model is binary, the error term is given by:

$$\epsilon_i = \begin{cases} 1 - x_i^T \beta, & \text{when } y_i = 1 \\ -x_i^T \beta, & \text{when } y_i = 0 \end{cases}$$  \hspace{1cm} (5)

This tells us that the error component cannot be normally distributed. Furthermore, the variance is not constant since

$$\text{Var}(y_i) = E[(y_i - E(y_i))^2] = (1 - \pi_i)^2 \cdot \pi_i + (0 - \pi_i)^2 \cdot (1 - \pi_i) = \pi_i \cdot (1 - \pi_i)$$  \hspace{1cm} (6)

is a function of the mean given by (4). Since $\pi_i$ only takes values between 0 and 1 and $x_i^T \beta$ is a real-valued function, the use of the linear model given in (1) can cause problems. Instead, we use the logistic response function which has the following form:

$$E(y_i) = \frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)} = \frac{1}{1 + \exp(-x_i^T \beta)}$$  \hspace{1cm} (7)

The logistic response function can be linearized by defining the linear regressor:

$$\eta = x_i^T \beta = \ln \frac{\pi}{1 + \pi}$$  \hspace{1cm} (8)

where $\eta$ is the logit transformation. [14] p. 423-424]

2.2.3. Logistic Regression Model

Thus, the general logistic regression model is given by:

$$y_i = E(y_i) + \epsilon, i = 1, \ldots, n$$  \hspace{1cm} (9)

where the response variable $y_i$ is an independent Bernoulli random variable with mean given in (7). [14] p. 423-424]
2.2.4. Maximum Likelihood

The Maximum Likelihood is a method for parameter estimation used in logistic regression. The likelihood function is given by:

\[ L(y_1, y_2, ..., y_n, \beta) = \prod_{i=1}^{n} f_i(y_i) = \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1-y_i} \]  

(10)

where the observations are assumed to be independent of each other. However, it is more practical to work with the logarithm of the likelihood function, or the log-likelihood:

\[ \ln L(y_1, y_2, ..., y_n, \beta) = \ln \prod_{i=1}^{n} f_i(y_i) = \sum_{i=1}^{n} y_i \ln \left( \frac{\pi_i}{1 - \pi_i} \right) + \sum_{i=1}^{n} \ln(1 - \pi_i) \]  

(11)

Maximizing the log-likelihood is the same as maximizing the likelihood. The maximum with respect to \( \beta, \hat{\beta} \), gives the vector of estimated regression coefficients. This is given by taking the derivative of the function, setting it to zero and solving it. [14, p. 425]

2.3. Model Evaluation and Diagnostics

2.3.1. Multicollinearity

Multicollinearity can be defined as a near-linear dependency among regressors. Formally, multicollinearity is present when we have \( t_1, t_2, ..., t_p \) as constants, not all zero, \( X_j \) is the \( j \):th column of the regressor matrix \( X \), and

\[ \sum_{j=1}^{p} t_j X_j = 0 \]  

(12)

holds for any subset of the columns of \( X \). However, even when the equation holds approximately, this dependency can cause a serious problem. The source of multicollinearity is often one of the following [14, p. 285-288]:

1. The data collection method employed
2. Constraints on the model or in the population
3. Model specification
4. An over-defined model

A commonly used statistic for determining multicollinearity is the Variation Inflation Factor, or VIF. The VIF is defined as:

\[ VIF = \frac{1}{1 - R_j^2} \]  

(13)
where $R_j^2$ is the coefficient of determination obtained when $x_j$ is regressed on the remaining regressors. It is usually considered a problem if the VIF is larger than 5 or 10. [14, p. 296]

Another way of checking multicollinearity is to compute the pairwise correlation matrix, that measures the correlation between each pair of regressors. A correlation close to unity implies that multicollinearity is present. [14, p. 289]

### 2.3.2. Akaike Information Criterion

The Akaike Information Criterion (AIC) is a measurement of the quality of a model. It is often used as a model evaluation criterion when comparing different models with each other and the model with the lowest AIC is preferred. Generally, AIC is defined as:

$$AIC = -2ln(L) + 2p$$

where $L$ is the likelihood function for a given model and $p$ is the number of parameters in the model. [14, p. 336]

### 2.3.3. Bayesian Information Criterion

The Bayesian Information Criterion (BIC) is, like the AIC, a measurement of the quality of a model. Like the AIC, the model with the lowest BIC is preferred. We define BIC as:

$$BIC = -2ln(L) + p \cdot ln(n)$$

where $L$ is the likelihood function for a given model, $p$ is the number of parameters and $n$ is the sample size. [14, p. 336]

### 2.3.4. Hypothesis Testing

A statistical hypothesis testing is constructed by stating a null hypothesis, $H_0$, and an alternative hypothesis, $H_1$. The null hypothesis is then tested for rejection given a significance level. In this project we will test whether each coefficient in our model is 0 or not. Each hypothesis is stated as [14, p. 436]:

$$H_0 : \beta_j = 0 \quad H_1 : \beta_j \neq 0$$
2.3.5. Wald Inference

One approach for performing hypothesis testing on the model coefficients is to use Wald inference. The Wald inference compares the maximum-likelihood estimator, \( \hat{\beta}_j \), to its standard error, \( se(\hat{\beta}_j) \). The test is defined as:

\[
Z_0 = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}
\]

For large samples the maximum-likelihood estimator is approximately normally distributed. The standard error is obtained by taking the square root of the covariance matrix of the estimated coefficients. The covariance matrix is given by:

\[
Var(\hat{\beta}) = -G \cdot (\hat{\beta})^{-1} = (X^TVX)^{-1}
\]

where \( G \) is the Hessian matrix. \( H_0 \) is rejected if \( P(\chi^2_\alpha > |z|) < \alpha \), where \( \alpha \) is the significance level. [14, p. 436-437]

2.3.6. Confidence Interval Estimation

Another way of checking the significance of the coefficients is to calculate the corresponding confidence interval for each coefficient. The confidence interval is based on the Wald inference and the 100(1-\( \alpha/2 \))% confidence interval is stated as:

\[
\hat{\beta}_j \pm z_{1-\alpha/2} \cdot se(\hat{\beta}_j)
\]

where 1-\( \alpha \) is the confidence level. The confidence level is the proportion of confidence intervals that contains the true value of the estimated coefficient. [10] p. 17-18

2.3.7. Cook’s Distance

The Cook’s distance is a tool to detect influential observations. It takes into account both the distance in x-space relative to the other observations, and the residual deviance [14, p.215]. In logistic regression, the measure is defined as [13]:

\[
C_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})^T X^TVX(\hat{\beta}_{(i)} - \hat{\beta})}{p}
\]

where \( \hat{\beta}_{(i)} \) is the maximum-likelihood estimated vector of coefficients excluding the i:th observation and \( \hat{\beta} \) is the coefficient vector including all observations. \( V \) is an \( n \times n \) diagonal matrix containing the variances for each observation. Since the observations are assumed to be uncorrelated, the off-diagonal elements are 0 [14, p.188].
2.3.8. ROC Curve

A way to validate the predictive ability of the generated model is through analyzing a Receiver Operating Characteristics (ROC) curve. The ROC curve plots the sensitivity of the model, the false negative rate, against the specificity, the true positive rate. The area under this curve is the probability for receiving a true positive for a range of different success probabilities. This is a measure of the model’s ability to discriminate a certain data set based on the outcome. [10 p. 160-161]

As a general rule, an Area Under Curve (AUC) value of 0.5 is as good as using a random model, that is, the model is of no use. A value between 0.7 and 0.8 implies ”acceptable discrimination”, while a value between 0.8 and 0.9 implies ”excellent discrimination”, and a value over 0.9 very rarely occurs in a real-world data set. [10 p. 162]

2.4. Variable Selection

2.4.1. Stepwise Regression

Stepwise regression is an approach based on various methods to obtain subset regression models. This is done by adding or removing regressor variables from the model based on certain model evaluation criteria. Two methods for performing stepwise regression are backward elimination and forward selection.

Backward elimination begins with the full model and then removes one variable at a time. The variable that is removed contributes the least to the model and the decision is based on different criteria such as AIC or BIC. Unlike the backward elimination, forward selection begins with no regressors in the model (intercept model) and then adds one variable at a time. [14 p. 344-347]

2.4.2. Lasso Regression

Lasso regression is a biased regression technique. Unlike maximum likelihood, which gives an unbiased estimation of the parameter coefficients, lasso regression generates estimations that have a slight bias, but with a lower variance. The bias term shrinks the coefficient values towards zero. Additionally, it sets some coefficients to exactly zero, thus performing variable selection. The quantity that lasso regression seeks to minimize is:

$$
\sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p}|\beta_j| \tag{21}
$$

The $\lambda$ tuning parameter can be selected arbitrarily, but is often determined with the aid of k-fold cross validation. When $\lambda$ is set to zero, the ordinary least squares estimation is obtained. [7 p. 219-220]
This is often presented as a minimization problem with a constraint on the length of the $\beta$-vector, that is:

$$\minimize_{\beta} \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

subject to $\sum_{j=1}^{p} |\beta_j| \leq c$ \hspace{1cm} (22)

By setting different values on $c$ the constraint can then be relaxed, to show for which lengths of the $\beta$-vector certain variables are included.

### 2.4.3. k-fold Cross-Validation

k-fold Cross-Validation (CV) is a method of determining the optimal square error, based on dividing up the data into $k$ number of "folds". The fit is determined using the other $k-1$ folds, and the $k$:th fold is used as validation data. For each fold, this produces a mean square error. The cross-validation error is then given by the average of all these, that is:

$$MSE = \frac{1}{k} \sum_{i=1}^{k} MSE_i$$

This estimation technique can also be used to estimate the $\lambda$ parameter. For a grid of $\lambda$-values, a cross-validation estimation can be ran, each giving a different CV error. These errors can then be compared, and the optimal solution can be determined, taking into account the deviance and the variable number reduction. [7] p. 225-228

3. Methodology

The method used in this project was logistic regression. We analyzed whether an outclick had occurred as a binary response variable (0 and 1) where 1 could be seen as a click on a search result and 0 as not a click.

3.1. Description of Data

3.1.1. Data Collection

Our data sample was taken from Flygresor’s live AI server. The data collected was from a one-week time period (2018-03-28 to 2018-04-04). The sample consisted of 4,338,286 searches. After the filtration was made, based on our stated limitations in Section 1.4, we ended up with a sample of 418,946 observations containing 6,367 outclicks.

We also collected data from the period 2018-04-06 to 2018-04-11 in order to test our derived model. The sample contained 186,130 observations with 2,583 outclicks.

3.1.2. Data Management

For handling the data, we used a local SQLite database which we imported to using Python3 scripts. The scripts were used to select the relevant information, and refining the data to fit our scope and limitations. The data was then imported as a data frame to RStudio, where we performed our statistical analysis.

3.1.3. Response Variable

The choice of response variable came naturally since our aim was to measure the outclick probability. Our response variable was 1 if an outclick has occurred, and 0 if it has not. Consequently, the expected value of our response variable gives us the outclick probability.
3.1.4. Regressor Variables

Given a vast range of options for variables, we defined some basic criteria for the candidate regressor variables. They had to:

1. Be correlated with the outclick probability
2. Be available for both outclick and non-outclick observations
3. Be ordinal or numeric, at least in an aggregated state

Following this, our regressor variables in the full model were adults, children, outboundStops, inboundStops, daysAhead and flightAlternatives (see Section 1.3). Starting with a high number of candidate regressor variables, we deemed these to be the most relevant.

A summary of the variables is given in table 1.

Table 1: Summary of the variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Min.</th>
<th>1st Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Qu.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>daysAhead</td>
<td>7</td>
<td>32</td>
<td>74</td>
<td>79.33</td>
<td>107</td>
<td>384</td>
</tr>
<tr>
<td>children</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.4591</td>
<td>0</td>
<td>21</td>
</tr>
<tr>
<td>adults</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1.894</td>
<td>2</td>
<td>9</td>
</tr>
<tr>
<td>inboundStops</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.7987</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>outboundStops</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.8132</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>flightAlternatives</td>
<td>1</td>
<td>10</td>
<td>20</td>
<td>31.78</td>
<td>33</td>
<td>58,336</td>
</tr>
</tbody>
</table>

3.2. Model Creation

3.2.1. Multicollinearity

When working with many variables, multicollinearity is often present. To examine this, we performed a VIF analysis on both our full model and our reduced model. This was done by regressing one regressor on the other regressors. Since strong multicollinearity is present when the VIF values are higher than 10, we wanted to remove variables with such values one at a time. Between each removal new VIF values needs to be obtained.

In addition to VIF, which measures one regressor’s correlation with all others, we also found it of value to check the pairwise correlations, the correlation between each pair of regressors. We wanted the pairwise correlations not to be close to unity.
3.2.2. Variable Selection

In order to obtain the best subset model, we used forward selection and backward elimination. For each step, we compared the reduced model with the previous model and accepted the reduced model if the AIC and BIC had decreased. The algorithms which we based our stepwise regression on is presented by Zhang and utilizes the built-in functions of RStudio [19].

We performed lasso regression with the purpose of confirming that the correct variables were included in the final model. We created separate matrices for our regressor variables observations and our response variable observations, and scaled all observations to unit length. These matrices were used to create models for 56 different values of the tuning parameter $\lambda$, sorted from low to high. To determine the optimal value we used a cross validation technique, which drastically decreases the standard deviation of the binomial deviance. Here, we measured the deviance for each value. The value rendering the lowest deviance was selected, and the corresponding influence levels could be seen from the plot of the coefficient values with scaled observations. The variable with the least influence was then removed, and a new model was fitted using the same maximum-likelihood criterion as the full model. [9]

3.3. Model Evaluation

3.3.1. Residual Analysis

Since we are using such vast amounts of data, individual observations are unlikely to affect the fit substantially. To confirm this, we analyzed the Cook's distance for all observations and set the threshold to unity.

3.3.2. Wald Inference

We stated our hypothesis tests and used the Wald inference to test whether we can reject the null hypotheses. We chose the significance level to 1%.

3.3.3. Confidence Interval Estimation

The confidence interval estimation was constructed on every estimated coefficient and the chosen confidence level was set to 99%.

3.3.4. ROC Curve

To determine the predictive ability of the reduced model, we generated an ROC curve and measured the area under the curve. We did this in three steps. First, we generated probability predictions for each test data observation based on the reduced model. Then, we created an ROC curve through plotting the sensitivity of the model, represented by the actual responses, against the specificity, represented by the predictions. Finally, we measured the AUC. [10 p.160-162]
4. Results

4.1. Data Selection

In our data import, we encountered 25 empty observation rows. Due to our high amount of observations, we concluded that this would not affect our result significantly and hence removed these.

4.2. Full Model

The full model was fitted using all six original regressors and using a logit transformation. The estimate of the coefficients, the corresponding standard error and the Wald inference is presented in table 2.

Table 2: Summary

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Estimate</th>
<th>Standard error</th>
<th>z-value</th>
<th>Wald inference</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-2.7492151</td>
<td>0.0325648</td>
<td>-84.423</td>
<td>(&lt; 2 \cdot 10^{-16})</td>
<td>***</td>
</tr>
<tr>
<td>daysAhead</td>
<td>0.0015703</td>
<td>0.0002153</td>
<td>7.292</td>
<td>3.05 \cdot 10^{-13}</td>
<td>***</td>
</tr>
<tr>
<td>children</td>
<td>-0.0022056</td>
<td>0.0100249</td>
<td>-0.220</td>
<td>0.826</td>
<td></td>
</tr>
<tr>
<td>adults</td>
<td>0.0690112</td>
<td>0.0101856</td>
<td>6.775</td>
<td>1.24 \cdot 10^{-11}</td>
<td>***</td>
</tr>
<tr>
<td>inboundStops</td>
<td>-1.3277943</td>
<td>0.0562109</td>
<td>-23.622</td>
<td>(&lt; 2 \cdot 10^{-16})</td>
<td>***</td>
</tr>
<tr>
<td>outboundStops</td>
<td>-1.4775363</td>
<td>0.0565987</td>
<td>-26.105</td>
<td>(&lt; 2 \cdot 10^{-16})</td>
<td>***</td>
</tr>
<tr>
<td>flightAlternatives</td>
<td>-0.0367425</td>
<td>0.0013407</td>
<td>-27.406</td>
<td>(&lt; 2 \cdot 10^{-16})</td>
<td>***</td>
</tr>
</tbody>
</table>

There seems to be a strong relationship between the response variable and all regressors but one, children. The Wald inference for this variable exceeds our chosen significance level of 1 %.

4.2.1. Multicollinearity

To determine whether the full model had a multicollinearity issue, we analyzed both the VIF values and the pairwise correlations. As can be seen in table 3, all VIF values are below 5.

Table 3: VIF

<table>
<thead>
<tr>
<th>Variable</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>daysAhead</td>
<td>1.066018</td>
</tr>
<tr>
<td>children</td>
<td>1.029194</td>
</tr>
<tr>
<td>adults</td>
<td>1.056475</td>
</tr>
<tr>
<td>inboundStops</td>
<td>1.898157</td>
</tr>
<tr>
<td>outboundStops</td>
<td>1.901841</td>
</tr>
<tr>
<td>flightAlternatives</td>
<td>1.091230</td>
</tr>
</tbody>
</table>

The pairwise correlation matrix was computed and is presented in figure 2. We note that there is no correlation close to unity.
4.2.2. Confidence Interval Estimation

The 99\% confidence intervals for the coefficients in the full model is given in table 4.

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>0.5% (Lower)</th>
<th>99.5% (Upper)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-2.833096407</td>
<td>-2.665333772</td>
</tr>
<tr>
<td>daysAhead</td>
<td>0.001015598</td>
<td>0.002124946</td>
</tr>
<tr>
<td>children</td>
<td>-0.028027934</td>
<td>0.023616689</td>
</tr>
<tr>
<td>adults</td>
<td>0.042774768</td>
<td>0.095247655</td>
</tr>
<tr>
<td>inboundStops</td>
<td>-1.472583993</td>
<td>-1.183004677</td>
</tr>
<tr>
<td>outboundStops</td>
<td>-1.623324765</td>
<td>-1.331747794</td>
</tr>
<tr>
<td>flightAlternatives</td>
<td>-0.040195805</td>
<td>-0.033289161</td>
</tr>
</tbody>
</table>

The result shows that the confidence interval for the coefficient \textit{children} does in fact include the value 0. This supports the result from our hypothesis testing.
4.2.3. **Cook’s Distance**

We performed residual analysis through calculating the Cook’s distances for each observation. Figure 3 shows the result. The maximum value was 0.2473.

![Figure 3: Cook’s distance](image)

Since the largest value is lower than 1, no observations were removed.

4.3. **Variable Selection**

4.3.1. **Forward Selection and Backward Elimination**

Forward selection and backward elimination was performed in order to obtain a model improvement of the full model. The criteria used were the AIC and BIC. Table 5 shows that forward selection recommends not to remove any variables from the full model, while the backward elimination method recommends to remove the variable *children*.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Forward Selection</th>
<th>Backward Elimination</th>
</tr>
</thead>
<tbody>
<tr>
<td>daysAhead</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>children</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>adults</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>inboundStops</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>outboundStops</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>flightAlternatives</td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>

Table 5: Stepwise regression
4.3.2. Lasso Regression with k-fold Cross Validation

We performed a lasso regression analysis for 56 different values of $\lambda$ with our unit scaled matrices. The most appropriate value of $\lambda$ was then selected using 10-fold cross validation, measuring the binomial deviance for each value, see figure 4. The first dotted vertical line shows the value of $\lambda$ that gives the lowest mean deviance.

Figure 4: Bias effect on variables and binomial deviance

The function is strictly increasing, meaning that even the smallest bias decreases the accuracy of the model. Therefore, no bias is used in the model. The bias analysis also shows that even an unbiased ordinary least-squares estimation of $\beta$ would contain five variables, thus causing the variable *children* to be removed.

4.3.3. Variable Influence

The relative influences on the fit using unit scaling can be seen in figure 5. The influences are relative to the standardized lengths of each regressor, meaning that regressors with a larger mean value will have a larger relative influence than in the unscaled model.

Figure 5: Variable influences on fit
4.4. Reduced Model

A reduced model was fitted using five regressors, daysAhead, adults, inboundStops, outboundStops and flightAlternatives. The estimate of the coefficients, the corresponding standard error and the Wald inference is presented in table 6.

Table 6: Summary

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Estimate</th>
<th>Standard error</th>
<th>z-value</th>
<th>Wald inference</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-2.7496283</td>
<td>0.0325163</td>
<td>-84.561</td>
<td>&lt;2 \cdot 10^{-16}</td>
<td>***</td>
</tr>
<tr>
<td>daysAhead</td>
<td>0.0015647</td>
<td>0.0002139</td>
<td>7.316</td>
<td>2.56 \cdot 10^{-13}</td>
<td>***</td>
</tr>
<tr>
<td>adults</td>
<td>0.0688139</td>
<td>0.0101498</td>
<td>6.780</td>
<td>1.20 \cdot 10^{-11}</td>
<td>***</td>
</tr>
<tr>
<td>inboundStops</td>
<td>-1.3276850</td>
<td>0.0562113</td>
<td>-23.620</td>
<td>&lt;2 \cdot 10^{-16}</td>
<td>***</td>
</tr>
<tr>
<td>outboundStops</td>
<td>-1.4775043</td>
<td>0.0566011</td>
<td>-26.104</td>
<td>&lt;2 \cdot 10^{-16}</td>
<td>***</td>
</tr>
<tr>
<td>flightAlternatives</td>
<td>-0.03607363</td>
<td>0.0013403</td>
<td>-27.408</td>
<td>&lt;2 \cdot 10^{-16}</td>
<td>***</td>
</tr>
</tbody>
</table>

Significance codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

We notice that all included regressors have a high significance, meaning that the null hypotheses cannot be rejected. Therefore, no further reduction of the model is necessary.

4.4.1. Multicollinearity

The VIF value for each corresponding variable is presented in table 7. We notice that all VIF values are below 5, indicating no presence of multicollinearity.

Table 7: VIF

<table>
<thead>
<tr>
<th>Variable</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>daysAhead</td>
<td>1.051246</td>
</tr>
<tr>
<td>adults</td>
<td>1.048329</td>
</tr>
<tr>
<td>inboundStops</td>
<td>1.898221</td>
</tr>
<tr>
<td>outboundStops</td>
<td>1.902036</td>
</tr>
<tr>
<td>flightAlternatives</td>
<td>1.090746</td>
</tr>
</tbody>
</table>
4.4.2. Confidence Interval Estimation

Table 8 shows the 99 % confidence intervals for each coefficient in the reduced model.

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>0.5 % (Lower)</th>
<th>99.5 % (Upper)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-2.833384858</td>
<td>-2.665871744</td>
</tr>
<tr>
<td>daysAhead</td>
<td>0.001013782</td>
<td>0.002115575</td>
</tr>
<tr>
<td>adults</td>
<td>0.042669685</td>
<td>0.094958095</td>
</tr>
<tr>
<td>inboundStops</td>
<td>-1.472475678</td>
<td>-1.182894316</td>
</tr>
<tr>
<td>outboundStops</td>
<td>-1.623299098</td>
<td>-1.331709588</td>
</tr>
<tr>
<td>flightAlternatives</td>
<td>-0.040188874</td>
<td>-0.033283807</td>
</tr>
</tbody>
</table>

4.4.3. Cook’s Distance

We calculated the Cook’s distances for each observation in the reduced model. The maximum value was 0.04947. The result is presented in figure 6.

![Figure 6: Cook’s distance](image)
4.5. Model Evaluation

4.5.1. ROC Curve and AUC

We plotted our ROC curve, as can be seen in figure 7. Then, we calculated the AUC, using the testing data, and received a value of 0.8586. According to Hosmer and Lemeshow, the model is considered to have "excellent discrimination", implying excellent predictive ability [10 p.162].

Figure 7: ROC curve
5. Discussion

5.1. Model Creation

5.1.1. Choice of Regression Model

For the purpose of answering our research question, there were several regression techniques that could have been used. The primary techniques we considered were multiple linear regression, poisson regression and logistic regression.

Multiple linear regression with other transformations than the logit transformation could have been considered. Given the right transformations, we could have obtained a model with the conversion rate as a response variable. Practically, however, this would be difficult, and the model achieved would have been unreliable.

Given a binary response variable, poisson regression could also have been a good tool for analyzing the conversion rate, especially since the rate was low, and the searches had a clear time dimension. We can recommend this method for further research. However, we are more familiar with logistic regression and it is a perfectly appropriate method to use. Logistic regression also has good support in R, making the analysis accurate and simple to represent.

On another note, simple statistical analysis of the data without producing a model was also considered. With that, we would have lost our predictive ability, and thus we rejected the idea.

5.1.2. Candidate Variables

We selected candidate variables from the three criteria defined in Section 3.1.4. The first criterion sorted out many candidates, as we had irrelevant variables available for this analysis. Some examples of these are what archive the searches were listed in, or the IP-addresses of the searches. The second criterion caused us not to select our flight price and travel time variable options. Since these were only available for searches with an outclick, they would not be meaningful for our analysis. An option discussed here was to put a pseudo-price and travel time for searches not resulting in an outclick through manipulating the data. Determining this would be very difficult since the users see a wide variety of prices and travel times for different destinations and different flights. Generating these would therefore give poor statistical significance. The second criterion along with the third criterion caused us to remove candidacy for the OTA and airline variables. Pseudo-values for these could be generated with the same poor accuracy. Additionally, these variables have no ordinance or numerical value. Flygresor offered to provide each OTA’s and airline’s recorded conversion rate, allowing us to put a numerical value on them. However, this alone was not enough to meet our requirements, and we left these variables out of the analysis.
5.1.3. Reduced Model

The result from the backward elimination, Wald inference and lasso regression supported the hypothesis that the variable *children* did not explain much of the outclick probability. Furthermore, when calculating the confidence interval for this variable, we noticed that the interval contained the value 0. Because various criteria and methods pointed to the same conclusions, we chose to remove the variable. Also, the decrease in the AIC and BIC and the low Wald inference for the reduced model makes removing additional variables unnecessary. Using the full model might produce an adequate model for simple analysis, but we deemed the reduced model to be of higher value since all its variables are significant. Moreover, a model with fewer variables is easier to use.

5.2. Model Accuracy

5.2.1. Cook’s Distance

Our main challenge here was to determine an appropriate threshold for removal. Different recommendations for what is considered a high value exist, where many aspects have to be taken into account. The Cook’s distances are generally lower on average in a large sample. In logistic regression, given a low-odds ratio, the values are generally higher for the ”success” outcomes. If these high values are removed, the results will not be representative of the data, since the removal significantly shifts the odds. We therefore chose Montgomery’s recommendation, removing values larger than unity [14, p.216].

5.2.2. Predictive Ability

The predictive ability of the model was examined through analyzing the ROC curve. The result from figure 7 showed an excellent predictive ability, and thus we can consider the model fit to use for Flygresor.

Given the excellent Wald inference statistics, however, one can question why the predictive ability is not higher for the model. This may be attributed to the high difference in behavior of individual customers or the low outclick ratio. The difference in behavior causes the predictive ability to be worse than the Wald inference might show; overall, behavior can be determined, but each individual might act in a very different way than the aggregated whole. The low outclick ratio can worsen this problem, since a given combination of favorable variable values still does not guarantee an outclick. The number of true positives is therefore low, and our sensibility values are decreased. This is, however, a given circumstance, and the predictive ability should be judged with this in mind. Our original recommendation therefore holds.
5.3. Model Interpretation

In our final model, we were left with five variables, each with a coefficient significantly deviating from zero. We will discuss the interpretation of these variables and their corresponding coefficients.

5.3.1. Variables and Coefficients

-daysAhead: this variable has a positive coefficient, meaning that a search made many days ahead of the departure will have a higher outclick probability than a search made shortly before departure. We interpret this as a result of the customer’s speculative ability. Information from Supersaver reveals that the conversion ratio from an outclick to an actual purchase decreases as daysAhead increases. We therefore conclude that these outclicks are primarily speculative; the customer does more thorough research on a trip being planned far ahead in time.

-adults: this variable also has a positive coefficient, meaning that a search made with many adults has a higher outclick probability than a search with few adults. We note that the standard setting for a search is one adult. Even if the intention is to travel with several adults, the prices can be manually multiplied by the customer when an appropriate trip is found. These searches are likely to be speculative, and when a customer has a higher interest in a purchase, a full search with all traveling members is performed.

-outboundStops/inboundStops: as can be seen in table 3 and figure 2, these two variables are correlated. This has an intuitive interpretation; booking a trip to an inaccessible city or airport requires many stops, and therefore the return journey requires the same amount of stops. Since the scope of this thesis is limited to return flights without an open-jaw option, this relation holds for most searches. For these variables, we have relatively large negative coefficients, meaning that a trip with many stops has a much lower outclick probability than a trip with no or few stops. We assume here that this is because of inconvenience; the higher the number of stops, the longer the travel time. Since time is valuable, people choose faster flights rather than longer ones.

-flightAlternatives: since the variable flightAlternatives has a negative estimated coefficient, an increase in the number of searches displayed for the user will decrease the probability of an outclick. The default number of searches displayed for the user is 10, provided that the user has not made any heavy selective filtrations. The number of searches displayed increases as the user scrolls down on the page. However, this does not mean that a trip with a low number of search results increases the probability of an outclick. Our interpretation of this variable is rather that the outclick probability decreases if the user scrolls down on the page. This shows that Flygresor’s first displayed results represent the customer’s demand well.
As can be seen in figure 5, the variable \textit{daysAhead} has less influence on the fit. It is known that people are more likely to click out on a flight far away from the current date, e.g. for a summer or Christmas break [1]. This causes the coefficient to be much lower. The \textit{adults} variable also has a low influence on the fit. This shows that the group size does affect the probability of an outclick, but not much.

The \textit{outboundStops}, \textit{inboundStops} and \textit{flightAlternatives} coefficient are more interesting. They all have a large negative coefficient. We could assume \textit{inboundStops} and \textit{outboundStops} to have large coefficients, as it is a well-known fact that customers generally choose faster and more convenient flights. We notice however that the coefficient for \textit{outboundStops} has a greater magnitude than the one for \textit{inboundStops}, meaning that customers would rather have fewer stops (and implicitly, shorter travel time) on the way out than the way in. The \textit{flightAlternatives} coefficient is an interesting finding. Being the most influential factor by far, it shows that more than the half of outclick decisions are based primarily on the number of alternatives given. This should therefore be the most important point to consider in decision making for Flygresor.

5.4. Model Implication

Since the variables \textit{outboundStops} and \textit{inboundStops} have the greatest effect on the probability of an outclick the most, given a unit change in the variable value, a recommendation to Flygresor is to decrease the number of flight alternatives with a high number of stops. Considering that \textit{outboundStops} has a slightly higher influence on the outclick probability than \textit{inboundStops}, another recommendation is to offer flight alternatives with fewer stops on the outbound trip than the inbound trip.

Furthermore, the result shows that with the normalization of our data sample, the variable \textit{flightAlternatives} has the highest influence on the outclick probability. Our recommendation for this is not to decrease the number of search results displayed for the user initially, but rather to increase the use of filtration on the site. A decrease in the number of search results would increase the risk of removing flight alternatives which would lead to an outclick. We therefore recommend Flygresor to promote the usage of filtration.

Although the variable \textit{daysAhead} has a low influence on the outclick probability, we recommend Flygresor to market flights around holidays and school breaks early.
5.5. Model Application

5.5.1. Usage Guidelines

The purpose of this regression model is to analyze the expected probability of an outclick. Note that this model generates approximate probabilities, and not exact values. Though the model has a high predictive ability, it does not guarantee that a predicted success (an outclick) will generate a success, or that a predicted failure (not an outclick) will generate a failure. The probability of such errors is given by the ROC curve in figure 7.

The model is intended for Flygresor’s usage, and though large parts of the result might be accurate for the industry as a whole, this is not guaranteed. The model is robust for internal variances during a week, but is not robust regarding seasonal variations. To ensure seasonal stability, testing has to be done with data samples collected over a longer period of time, i.e. at least a year. As stated in Section 1.4, the model is made for trips within Europe, excluding the Nordic countries. It is made to determine leisure traveller behavior, not business traveller behavior.

In the model, values for all variables are mandatory. To do a sensitivity analysis on one variable, all other variables need to remain constant. The variables included in the model do not fully explain customer behavior; there could be many other factors affecting the conversion rate. Finally, note that the outclick probability is uncorrelated with the probability of an actual flight purchase, and has no predictive ability for such observations.

5.6. Further development

Though our model describes the outclick probability with some accuracy, the model can be improved in several ways. For the purpose of answering our research question, it is worth doing an analysis of other factors not included in our model. The two primary factors that could have an influence are the ticket price and the travel time. To include those, a standard has to be set for when an outclick is not made. Preferably, these should include some median or mean price and travel time values, possibly with a transformation. There could also be factors affecting the probability not included in the original data, such as weather or customer age. These might be found through thorough user testing, and can then be measured and included in an analysis.

The model should also be made geographically and seasonally robust, if possible. To do this, the model needs to be tested and adjusted with data from different regions and over a full year, to see if it holds for such variations. If not, several models are needed to accurately describe reality. Finally, the model might also be expanded to describe the flight comparison industry as a whole, using data from other comparison sites, nationally or globally.
6. Conclusions

The purpose of this study was to analyze which factors affect the conversion rate at the flight comparison website flygresor.se the most. Starting with the factors presented in Section 1.3, the result showed that the factors that affect the conversion rate are:

- Number of adults (adults)
- Number of stops on the outbound trip (outboundStops)
- Number of stops on the inbound trip (inboundStops)
- Number of days between the search date and the departure date (daysAhead)
- Number of search results displayed for the user (flightAlternatives)

The factor with the highest influence on the outclick probability, given a unit change in the variable value, is outboundStops, followed by inboundStops, adults, flightAlternatives and daysAhead. However, daysAhead and flightAlternatives have high mean values. With normalization, the influence order is flightAlternatives, outboundStops, inboundStops, daysAhead and adults.

The final model is given by:

\[ \hat{y} = \frac{\exp(x^T \hat{\beta})}{1 + \exp(x^T \hat{\beta})} = \frac{1}{1 + \exp(-x^T \hat{\beta})} \]

where \( \hat{y} \) is the estimated outclick probability, \( x^T = (1, \text{daysAhead}, \text{adults}, \text{inboundStops}, \text{outboundStops}, \text{flightAlternatives}) \) and \( \hat{\beta}^T = (-2.7496283, 0.0015647, 0.0688139, -1.3276850, -1.4775043, -0.0367363) \) is the estimated vector of coefficient values.
References


