AGH UNIVERITY OF SCIENCE AND TECHNOLOGY
FACULTY OF APPLIED MATHEMATICS

The Selection of a Model for Airlines Customer Satisfaction

Joanna Krężel, Anna Matysek, Piotr Mikler, Adam Szczerba

14 January, 2022

Abstract

The project aims to analyze the data about satisfaction of Investico Airlines passengers. Having customer-granular observations about the cruise and a reported satisfaction level for particular aspects of the flight we try to fit models which predict whether they are satisfied with the service or not. The binary classification models used during the project are {} {} {}, out of which our recommendation is {} based on {}. This document describes the process we undertook and presents the results of data preprocessing, model selection and model validation.

Contents

1 Introduction

Tbd...

2 Data

The data is downloaded from www.kaggle.com and delivered by an airline organization. The dataset consists of the details of customers who have already flown with them. The feedback of the customers on various context and their flight data has been consolidated. The main purpose of this dataset is to predict whether a future customer would be satisfied with their service given the details of the other parameters values. Also the airlines need to know on which aspect of the services offered by them have to be emphasized more to generate more satisfied customers. The data consists of 129880 rows and 23 columns.

Below we list all column names with explanations of the variables' meaning. For categorical variables describing satisfaction level, 0 means *Not Available* and reflects situation in which the passenger did not provide a rating.

Feature	Description	Values
Satisfaction	Airline satisfaction level	satisfied/dissatisfied
Gender	Gender of the passenger	male/female
Customer type	The customer type	loyal / disloyal customer
Age	The age of a passenger	[7; 85] years
Type of travel	Purpose of the flight	personal / business travel
Class	Travel class in the plane	business / eco / eco plus
Flight distance	The flight distance of the journey	[50; 6951] miles
Seat comfort	Satisfaction level of seat comfort	{0-5}
Departure/arrival	Satisfaction level of departure/arrival time	{0-5}
Food and drink	Satisfaction level of food and drink	{0-5}
Gate location	Satisfaction level of gate location	{0-5}

Feature	Description	Values
Inflight WiFi service	Satisfaction level of the inflight wifi service	{0-5}
Inflight entertainment	Satisfaction level of inflight entertainment	{0-5}
Online support	Satisfaction level of online support	{0-5}
Ease of online booking	Satisfaction level of online booking	{0-5}
On-board services	Satisfaction level of on-board service	{0-5}
Leg room service	Satisfaction level of leg room service	{0-5}
Baggage handling	Satisfaction level of baggage handling	{0-5}
Checkin service	Satisfaction level of check-in service	{0-5}
Cleanliness	Satisfaction level of cleanliness	{0-5}
Online boarding	Satisfaction level of online boarding	{0-5}
Departure delay in minutes	Delay upon departure	[0; 1592] minutes
Arrival delay in minutes	Delay upon arrival	[0; 1584] minutes

2.1 Exploratory Data Analysis

Describe the ideas of this section. To be done, not urgent...

2.2 Data Preprocessing

As both academia and business point out, the data-related operations typically constitute about 80% of the whole effort of a modeling pipeline. The performance of any model is heavily driven by the quality of it's inputs. It can be easily proven in a simple trial by combat that even a suboptimal model running on high quality data can oftentimes bring a sophisticated one with poor inputs to it's knees. For that reason it is of utmost importance to

pay extra care and attention to the data which is fed to the decision making models.

During data preprocessing step we will gain insight about data statistics and information it conveys. First, we'll deal with NAs and outliers. Then we will refactor the features with variable binning and select a subset of features which are likely to display predictive power for our problem. In the end we will encode the variables and prepare a train-test split for model development.

2.2.1 Missing data treatment

Table 2: NA breakdown per feature. NAs span a small portion of data

Feature	NA_count	pct_of_data
ScheduleNote	6664	5.13%
FoodNote	5945	4.58%
SeatNote	4797	3.69%
EntertainmentNote	2978	2.29%
LegRoomNote	444	0.34%
ArrivalDelay	393	0.3%
WifiNote	132	0.1%
eBookingNote	18	0.01%
eBoardingNote	14	0.01%
CleanNote	5	0%
ServiceNote	5	0%
GateNote	2	0%
CheckInNote	1	0%
eSupportNote	1	0%

After examining the data it seems we don't have any critical issue related to missing values. *NAs* are present in 12 variables, but they constitute a minuscule portion of a very large dataset (see fig. 2). We considered employing an imputation strategy based on median, but given that NAs constitute roughly 0.08 of all observations, even if we drop them we would still have 119255 observations left to work with. Based on that we decided not to introduce imputed values to the dataset, but rather work with pure

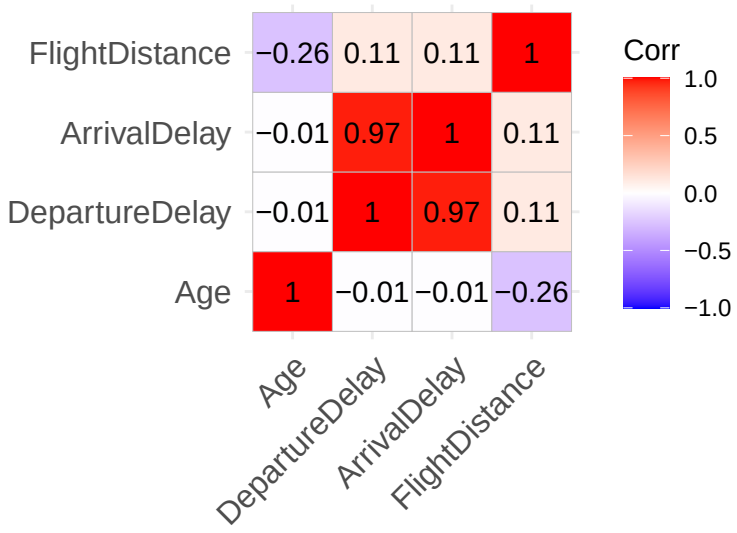
data.

2.2.2 Feature Engineering

The more is not always the better. Feature engineering is a pre-modeling stage which serves identifying features which are significant and filtering out the ones that are not. *Filtering feature selection methods* allow one to discard redundant features in a model independent way. By reducing the number of variables or discretizing them we simplify the model and increase it's interpretability. It is also a step which tackles multicollinearity (high linear codependency of explanatory variables) which kills stability and predictive power of some models.

Across the following sections we are going to introduce a few additional modifications of variables to the dataset and see if it makes sense to keep them. Since we have a big share of $0s$ in column `DepartureDelay`, we wanted to check if it makes sense to include a binary variable $IsDelayed$ as a predictor. We are also going to try discretizing `Age`, `DepartureDelay` and `FlightDistance` continuous variables into bins based on Weight of Evidence metric and evaluate their predictive power.

2.2.2.1 Continuous Features We have four continuous variables in our dataset: `Age`, `DepartureDelay`, `ArrivalDelay` and `FlightDistance`. We start the analysis by analyzing their codependence structure. We note a high linear relationship between `ArrivalDelay` and `DepartureDelay`, visible both in the correlation matrix (fig. **??**) and on figure 1. We can safely drop `ArrivalDelay`, since it doesn't introduce new information and additionally contaminates the dataset with NAs.



Next, we're going to examine the loess estimator of satisfaction as a function of the remaining continuous variables to see if any of them looks flat enough to raise suspicion regarding it's utility. Flatness of loess implies that average satisfaction does not change in the explanatory variable, hence the explanatory variable doesn't convey much information. In our case, this is not visible on the figure **??**, so we cannot discard any feature based on that. Note, the plot has been generated on a randomized data sub-sample for computational complexity reduction. We made sure however to take a sample large enough, so that the standard errors are tamed, and to ensure the relationship shape is stable regardless of the random seed chosen.

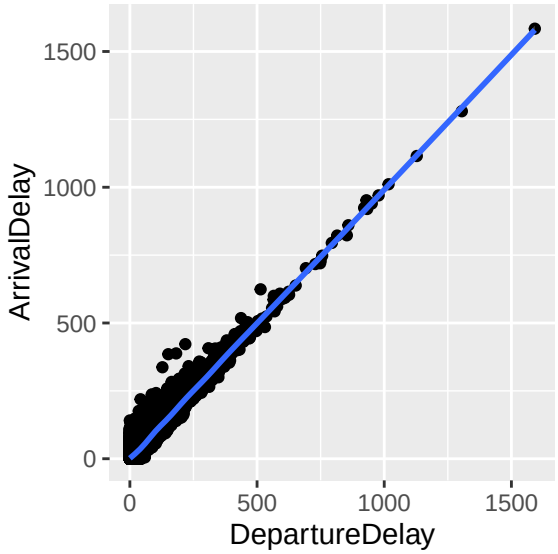
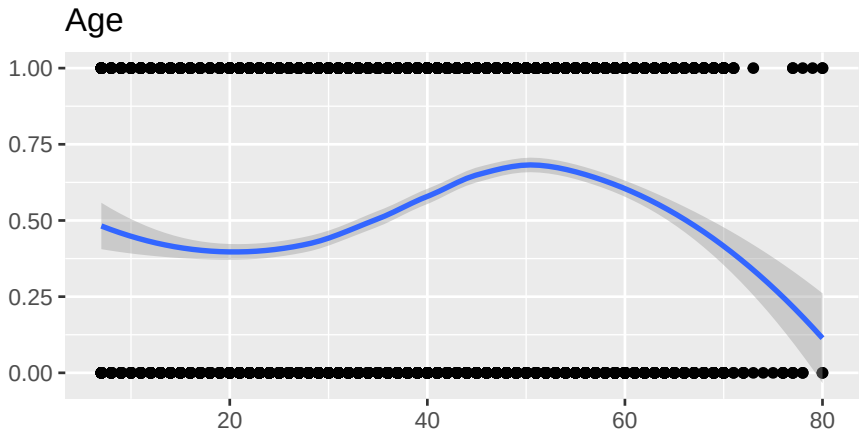
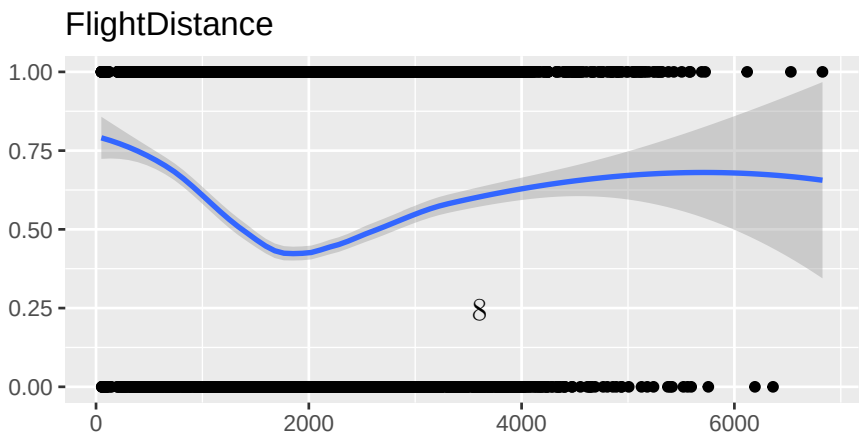


Figure 1: Strong linear relationship between departure delay and arrival delay allows to drop one of them from the dataset.





Now we are going to look at possible binnings of our features. We'll use `woeBinning::woe.binning` function which chooses the binning to maximize the information value of the feature. If the optimized binning will yield $IV < 0.1$, we will discard the variable. Otherwise we'll analyze the bins to ensure they are not over-optimized to an unreasonable degree. Based on loess plot shapes/regimes the expectation is to have not more than four bins for `Age` and `FlightDistance`, and a maximum of two bins for `DepartureDelay`.

WOE Table for Age

	Final.Bin	Total.Count	Total.Distr.	0.Rate	WOE	IV
1	<= 28	30896	25.9%	58.6%	-51.4	0.068
2	<= 40	29312	24.6%	49.0%	-12.6	0.004
3	<= 59	47672	40.0%	33.9%	50.1	0.096
4	<= Inf	11375	9.5%	53.2%	-29.3	0.008
6	Total	119255	100.0%	45.9%	NA	0.177

WOE Table for FlightDistance

	Final.Bin	Total.Count	Total.Distr.	0.Rate	WOE	IV
1	<= 1359	29839	25.0%	33.3%	53.0	0.067
2	<= 3053	71534	60.0%	53.4%	-30.2	0.055
3	<= Inf	17882	15.0%	36.7%	38.0	0.021
5	Total	119255	100.0%	45.9%	NA	0.143

WOE Table for DepartureDelay

	Final.Bin	Total.Count	Total.Distr.	0.Rate	WOE	IV
1	<= 19	95920	80.4%	44.0%	7.6	0.005
2	<= Inf	23335	19.6%	53.6%	-31.1	0.019
4	Total	119255	100.0%	45.9%	NA	0.024

From the WOE tables above we see that `DepartureDelay` is a variable of low predictive power, hence we won't use it in modelling. For the other variables, as the data binning chosen by the algorithm seems reasonable given

the ex-ante expectations, we're going to keep them.

2.2.2.2 Ordinal & Categorical Features The main challenge of the data preparation in this dataset is the proper treatment of passenger notes. Take for example the `SeatNote` feature which is a customer note describing their satisfaction level with the seating arrangement. One could ask himself the following questions:

- What did the passenger have in mind? Satisfaction with their seat location? Seat comfort? Possibility of choosing the seat?
- Does '$SeatNote$' = 3 imply a negative attitude towards a service? Or it's a moderate 'OK'?
- Is the satisfaction *"difference"* between notes 3 and 2 the same as between notes 5 and 4?
- Is a note '$SeatNote$' = 5 given '$Class$' = 'Eco' the same as '$SeatNote$' = 5 given '$Class$' = '$Business$'?

The same point is valid for any note-type variable in the dataset. The issue boils down to the problem that there is **no universal "unit" of satisfaction**. It is just as non-trivial to measure it as to predict it - simply because everyone perceives satisfaction in a subjective way. Our problem has an additional layer of complexity since we don't have information how precisely the survey questions have been described to the customers - so even if we *did* have some carefully designed satisfaction unit, we cannot be sure if all respondents referred to the same aspects of service when filling out the survey.

We aim to overcome this problem, by binning notes into wider classes, depending on how well they explain and affect the overall satisfaction. Since we have a lot of 5-leveled `Note` factors, there's a strong suspicion that in such large set there must exist some adjacent levels such that overall satisfaction is invariant to displacements in that group of levels. In other words, we could collapse notes of $1, 2\&3$ to one group if they carried similar information. Hence we will again let `woe.binning` automatically select bins and then verify the result.

Since there are 12 `Note` variables, we will only display one of the WOE tables as an example. However for all it has been verified that **adjacent** levels have been binned, so the binning is plausible, and the IV of the newly binned features are above 0.1.

Final.Bin	Total.Count	Total.Distr	1.Count	0.Count	1.Distr.	0.Distr.	0.Rate	WOE	IV
1	11713	9.8%	1561	10152	2.4%	18.6%	86.7%	-203.8	0.329
3 + 2	38010	31.9%	11063	26947	17.1%	49.3%	70.9%	-105.6	0.340
5 + 4	69532	58.3%	51945	17587	80.4%	32.2%	25.3%	91.7	0.443
Total	119255	100.0%	64569	54686	100.0%	100.0%	45.9%	NA	1.111

Next, we will check the information value for other categorical variables. They are binary, so binning has not been applied to them in the earlier step. The table 7 presents features and their IV. We see that there are features with IV smaller than 0.1 - we are going to drop those from the dataset.

Lastly, a brief look at the spearman rank correlation matrix (fig. 2) shows that there are no highly correlated "note" features among ordinal variables in the dataset.

Table 7: Information Value for all variables.

varName	IV
IsPersonalTravel	0.0532949
FlightDistance	0.1433763
Age	0.1769584
IsFemale	0.1893377
FoodNote	0.2330999
WifiNote	0.2920904
CheckInNote	0.3607163
Class	0.4001969
IsLoyal	0.4537176
CleanNote	0.4555277
BaggageNote	0.4718849
LegRoomNote	0.5714645
eBoardingNote	0.5948280
ServiceNote	0.6151573
eSupportNote	0.9223110
eBookingNote	1.1114520
SeatNote	1.4504018

varName	IV
EntertainmentNote	2.2947658

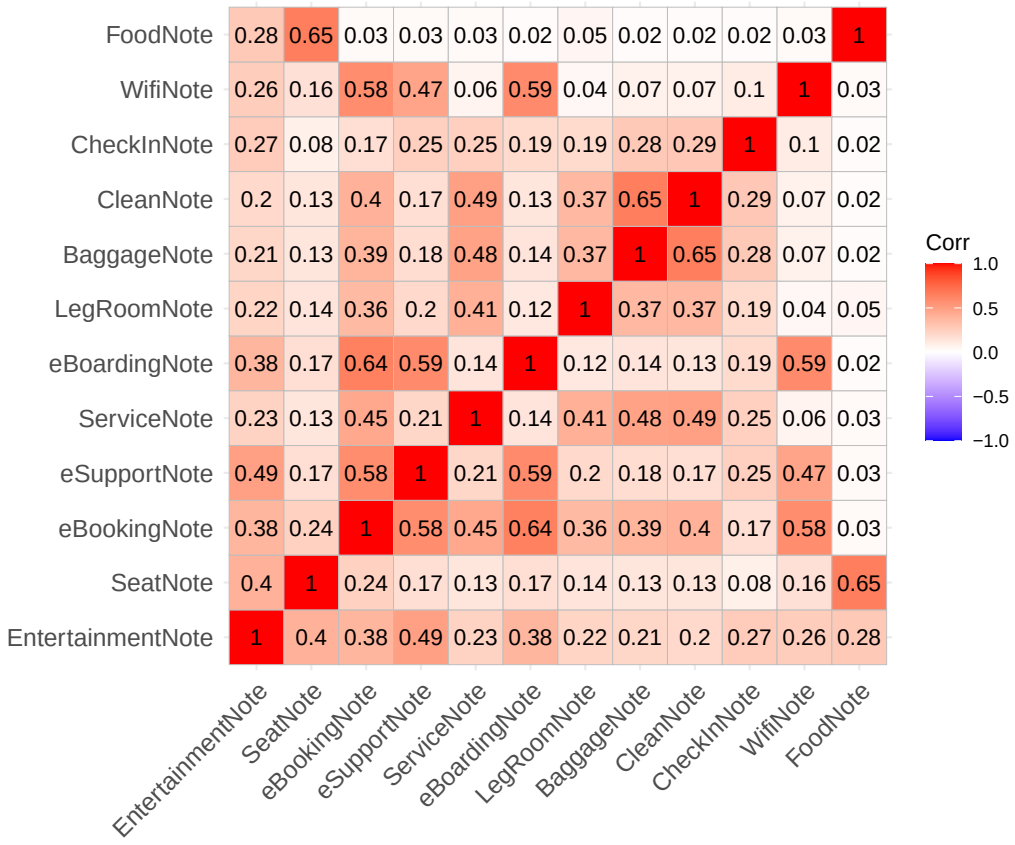


Figure 2: Spearman correlation shows no significant colinear relationships in ordinal variables

2.2.3 Feature Encoding

Some machine learning algorithms require numerical data, so we considered **ordinal encoding** and **dummy encoding** to transform our data.

We ran the following thought experiment to determine which encoding to employ. Say we use ordinal encoding and assign numbers to each factor

level. We could encode `Class` this way and assign a mapping like: {'Eco': 1, 'EcoPlus': 2, 'Business': 3}. This type of representation ensures the quality of the service is properly represented in the numeric data, but the question is whether this translates the same to the overall satisfaction? Yes, the business class is clearly more comfortable to travel in, but the *expectations* (the baseline) of business-class passengers will also be quite higher than the expectations of say, passengers in the economic class.

We chose to employ dummy encoding to encode `Class` - to avoid making assumptions about baseline satisfaction criteria across different passenger classes. For `Note` features however this problem is non-existent, since a higher note should correspond to higher satisfaction for any rational passenger. Here to avoid inflating the dataset with additional $4 \cdot 14 - 14 = 42$ sparse binary columns we will stick to the original ordinal encoding. This choice nonetheless should be revisited and controlled once we reach the stage of model choice and model fitting.

The resulting, encoded dataframe looks the following way:

```
## Rows: 119,255
## Columns: 34
## $ EntertainmentNote.M <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ EntertainmentNote.H <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ SeatNote.M          <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ SeatNote.H          <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ eBookingNote.H      <fct> 0, 1, 0, 0, 1, 1, 1, 0, 1, 0, 1, 1, 1, 0, 1, 0,...
## $ eBookingNote.M      <fct> 0, 0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 1,...
## $ eSupportNote.M      <fct> 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0,...
## $ eSupportNote.H      <fct> 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0,...
## $ ServiceNote.M       <fct> 0, 1, 1, 0, 0, 1, 1, 1, 1, 0, 0, 0, 1, 0, 0, 1,...
## $ ServiceNote.L       <fct> 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0,...
## $ eBoardingNote.H     <fct> 0, 1, 0, 0, 1, 1, 1, 0, 1, 0, 1, 0, 1, 1, 1, 0,...
## $ eBoardingNote.M     <fct> 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ LegRoomNote.M       <fct> 0, 1, 1, 0, 0, 1, 0, 0, 1, 0, 1, 0, 1, 1, 0, 0,...
## $ LegRoomNote.H       <fct> 0, 0, 0, 1, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 1,...
## $ BaggageNote.M       <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0,...
## $ BaggageNote.H       <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ CleanNote.M         <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0,...
## $ CleanNote.H         <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
```

```
## $ CheckInNote.L       <fct> 0, 0, 1, 1, 1, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0,...
## $ CheckInNote.H       <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ WifiNote.H          <fct> 0, 1, 0, 0, 1, 1, 1, 0, 1, 0, 1, 1, 1, 0, 1, 0,...
## $ WifiNote.M          <fct> 0, 0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 1,...
## $ FoodNote.M          <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ FoodNote.H          <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ Age.30s             <fct> 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0,...
## $ Age.40s50s          <fct> 1, 1, 0, 1, 0, 0, 1, 0, 0, 1, 1, 0, 1, 0, 0, 1,...
## $ Age.60plus          <fct> 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0,...
## $ FlightDistance.M    <fct> 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 0,...
## $ FlightDistance.H    <fct> 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1,...
## $ Class.Business      <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0,...
## $ Class.EcoPlus       <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1,...
## $ IsSatisfied         <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ IsFemale            <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ IsLoyal             <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...
```

3 A train-test split of data

4 Baseline Models

Before jumping straight into cutting-edge mathematical models, sometimes it can be very teaching to fit a simple models and analyze how well they perform on the data. It sets ground zero for any more complicated models that follow and allows to adjust own expectations. Therefore we will fit a simple perceptron on our train set and evaluate the performance on test set, to understand the complexity of our problem. ## Fitting and performance

Conceptually, perceptron consists of two simple parts: a linear combination of inputs, and the sigmoid activation function. Perceptron optimizes the weights of the components of the linear sum, to arrive at the best fit on the training set.

In our case, the fitted weights are presented in

x

<table class='gmisc_table' style='border-collapse: collapse; margin-top: 1em; margin-bottom: 1

	Perceptron_metrics
Accuracy	0.8893128
Sensitivity	0.8804129
Specificity	0.8968617
Precision	0.8786470
Recall	0.8804129

4.1 Validation

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

5 The Challenger Models

5.1 Random Forest

5.1.1 Fitting and performance

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

5.1.2 Validation

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

5.2 Logistic Regression

5.2.1 Fitting and performance

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

5.2.2 Validation

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

5.3 State-of-the-art: Neural Network model

5.3.1 Fitting and performance

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

5.3.2 Validation

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

6 Conclusion

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

7 Bibliography