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

10 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 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.

As the first step in our modeling pipeline we are going to look at the dataset

to gain insight about it's statistics and information it conveys. We'll refactor the feature names to something more manageable and represent accordingly different data types present in the dataset. Lastly we will perform quality checks on the data, such as outlier detection and treatment of not-available values (*NAs*).

2.1.1 Feature Encoding

Most machine learning algorithms require numerical inputs. Our data is mostly categorical and ordinal, hence we need to start with encoding those features.

2.1.1.1 Categorical features The dataset contains some binary categorical information such as *Male/Female*, *Loyal/Disloyal Customer*, etc. We are going to employ binary encoding for those features, that is: map values to 1 or 0 and rename the factors to `IsSatisfied`, `IsFemale`, `IsLoyal` for easier interpretation.

2.1.1.2 Ordinal features The main challenge of the data preparation in this dataset is the treatment of ordinal features. 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 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 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 it 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.

Before discussing this further let's take a short detour to the options we have when dealing with ordinal variables for Machine Learning. Two most common approaches emerge: **Dummy encoding** and **Ordinal encoding** - both are valid, depending on what we're trying to achieve.

We could use ordinal encoding and assign numbers to each vote. This is pretty much what we already have in our "note" features. 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. This may result in a counterintuitive drop in the satisfaction level, simply because the sub-populations across business classes will perceive the service differently.

The second possibility we have for encoding ordinal variables is the *dummy encoding* which will split the feature `Class` into features: `Class.Eco`, `Class.EcoPlus` and `Class.Business` assigning ones and zeros in appropriate places. One of those features will be dropped to avoid perfect linear relationship (otherwise the sum of the new features would always be 1), but we'll not lose information. We only need $n - 1$ features to encode full information about a factor with n possible levels.

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 $4 * 14 - 14 = 42$ 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: 129,880
## Columns: 24
```

```
## $ Class.Business    <fct> 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, ...
## $ Class.EcoPlus     <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ Age               <dbl> 65, 47, 15, 60, 70, 30, 66, 10...
## $ FlightDistance    <dbl> 265, 2464, 2138, 623, 354, 189...
## $ SeatNote          <fct> NA, NA, NA, NA, NA, NA, NA, NA...
## $ ScheduleNote      <fct> NA, NA, NA, NA, NA, NA, NA, NA...
## $ FoodNote          <fct> NA, NA, NA, NA, NA, NA, NA, NA...
## $ GateNote          <fct> 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, ...
## $ WifiNote          <fct> 2, NA, 2, 3, 4, 2, 2, 2, 5, 2,...
## $ EntertainmentNote <fct> 4, 2, NA, 4, 3, NA, 5, NA, 3, ...
## $ eSupportNote      <fct> 2, 2, 2, 3, 4, 2, 5, 2, 5, 2, ...
## $ eBookingNote      <fct> 3, 3, 2, 1, 2, 2, 5, 2, 4, 2, ...
## $ ServiceNote       <fct> 3, 4, 3, 1, 2, 5, 5, 3, 4, 2, ...
## $ LegRoomNote       <fct> NA, 4, 3, NA, NA, 4, NA, 3, NA...
## $ BaggageNote       <fct> 3, 4, 4, 1, 2, 5, 5, 4, 1, 5, ...
## $ CheckInNote       <fct> 5, 2, 4, 4, 4, 5, 5, 5, 5, 3, ...
## $ CleanNote         <fct> 3, 3, 4, 1, 2, 4, 5, 4, 4, 4, ...
## $ eBoardingNote     <fct> 2, 2, 2, 3, 5, 2, 3, 2, 4, 2, ...
## $ DepartureDelay    <dbl> 0, 310, 0, 0, 0, 0, 17, 0, 0, ...
## $ ArrivalDelay      <dbl> 0, 305, 0, 0, 0, 0, 15, 0, 0, ...
## $ IsSatisfied       <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
## $ IsFemale          <fct> 1, 0, 1, 1, 1, 0, 1, 0, 1, 0, ...
## $ IsLoyal           <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
## $ IsPersonalTravel  <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
```

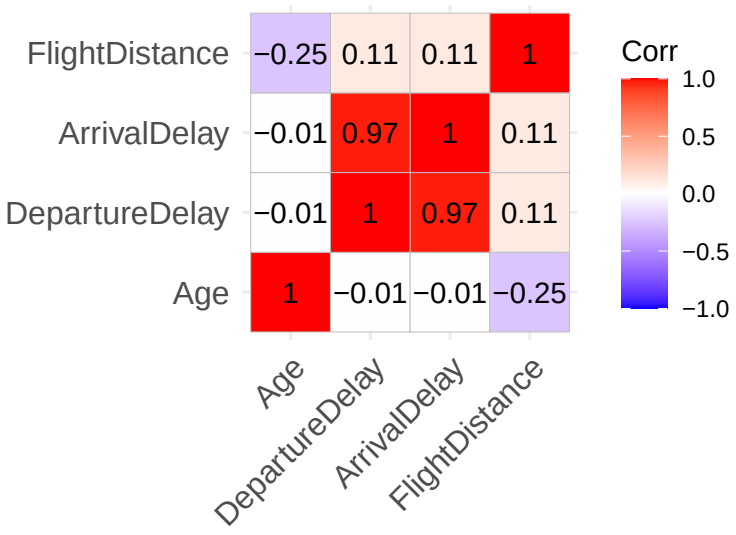
2.1.2 Feature Engineering

The more is not always the better. Every model has a certain computational complexity that increases with the number of additional explanatory variables. 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 they 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 0*s* 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.

2.1.2.1 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 (very heuristically) 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

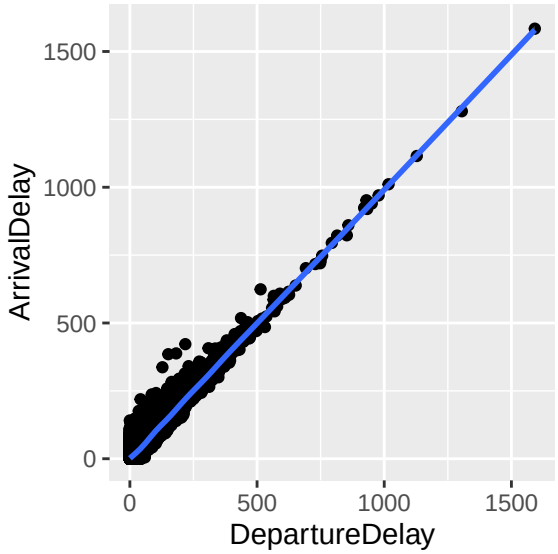
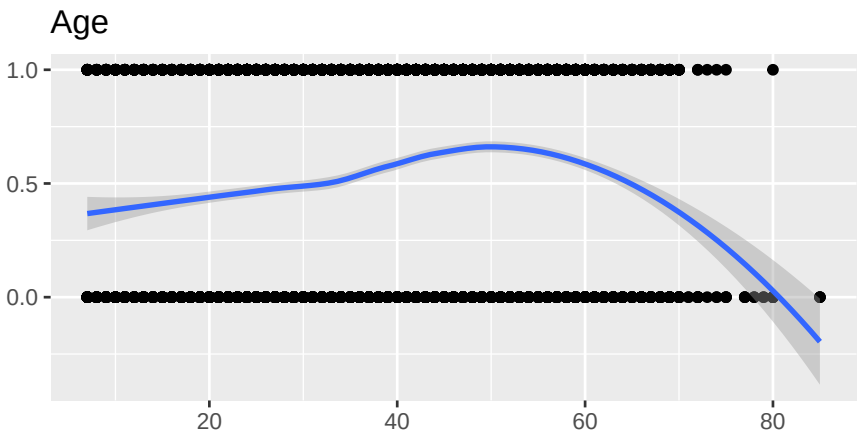
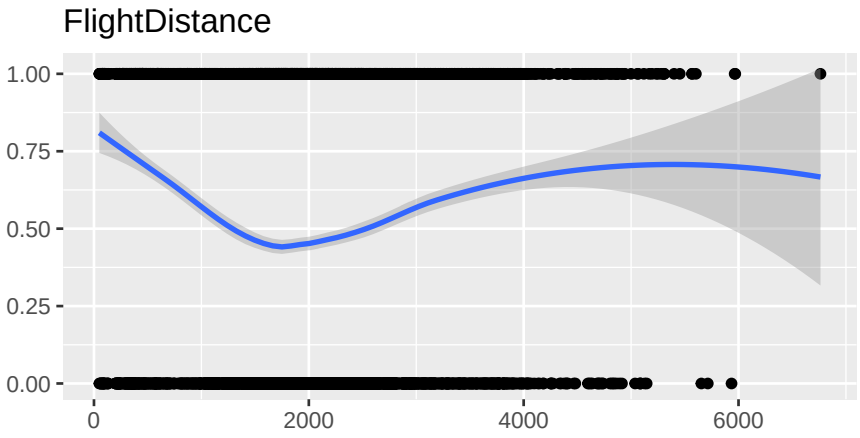
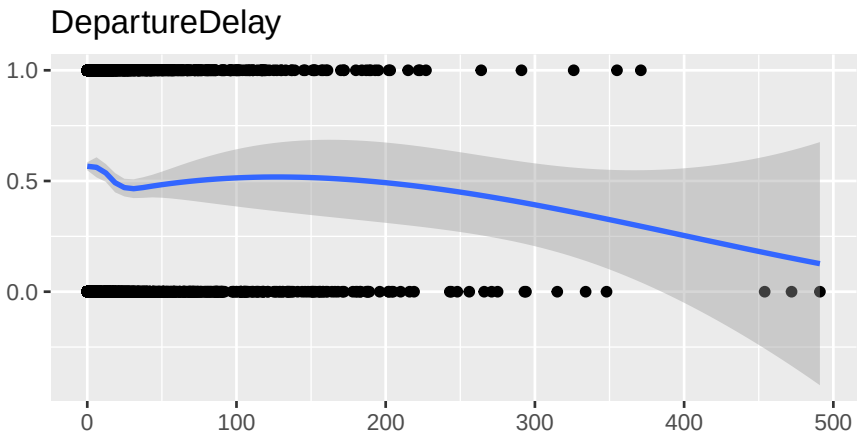


Figure 1: Strong linear relationship between departure delay and arrival delay allows to drop one of them from the dataset.

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.


Age
1.0
0.5
0.0
20
40
60
80


FlightDistance
1.00
0.75
0.50
0.25
0.00
0
2000
4000
6000


DepartureDelay
1.0
0.5
0.0
0
100
200
300
400
500

2.1.2.2 Categorical Features For categorical features we will use the chi-square test and information value to

- detect dependency between the response variable and feature in question
- quantify the predictive power of the categorical variables. It is customary to take information value smaller 0.1 to mean "weak predictive power", so we will take that as our cut-off.

The Chi Square test rejected independency hypothesis of all categorical variables and the target variable. However, the table 2 presents that there are features with IV smaller than 0.1. That means even though they are co-dependent with satisfaction, their predictive power is not big. 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 2: Categorical features of low informational value

varName	IV	pvalChSq
ScheduleNote	0.0068857	0
IsDelayed	0.0131859	0
IsPersonalTravel	0.0480765	0
GateNote	0.0831380	0

2.1.3 Missing data treatment

At this point, in the dataset we only have *NAs* corresponding to a missing passenger notes. Let's tackle that issue in this short section.

Table 3: NA breakdown per feature. NAs span a small portion of data

Feature	NA_count	pct_of_data
FoodNote	5945	4.58%
SeatNote	4797	3.69%
EntertainmentNote	2978	2.29%

Feature	NA_count	pct_of_data
LegRoomNote	444	0.34%
WifiNote	132	0.1%
eBookingNote	18	0.01%
eBoardingNote	14	0.01%
CleanNote	5	0%
ServiceNote	5	0%
CheckInNote	1	0%
eSupportNote	1	0%

Generally speaking we don't have any critical issue related to missing values in our data. Yes, there are *NAs* present in 12 variables, but they constitute a minuscule portion of a very large dataset (see fig. 3). We considered employing an imputation strategy based on median, but given that NAs constitute roughly 0.05 of all observations, even if we drop them we would still have 123554 observations left to work with. Based on that we decided not to introduce imputed values to the dataset, but rather work with pure data.

2.1.3.1 Variables binning 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 variable. 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.

WOE Table for FlightDistance

	Final.Bin	Total.Count	Total.Distr.	0.Rate	WOE	IV
1	<= 867	18555	15.0%	26.4%	88.7	0.108
2	<= 3406	92658	75.0%	51.9%	-21.7	0.036
3	<= Inf	12341	10.0%	36.1%	43.3	0.018
5	Total	123554	100.0%	46.5%	NA	0.161

WOE Table for Age

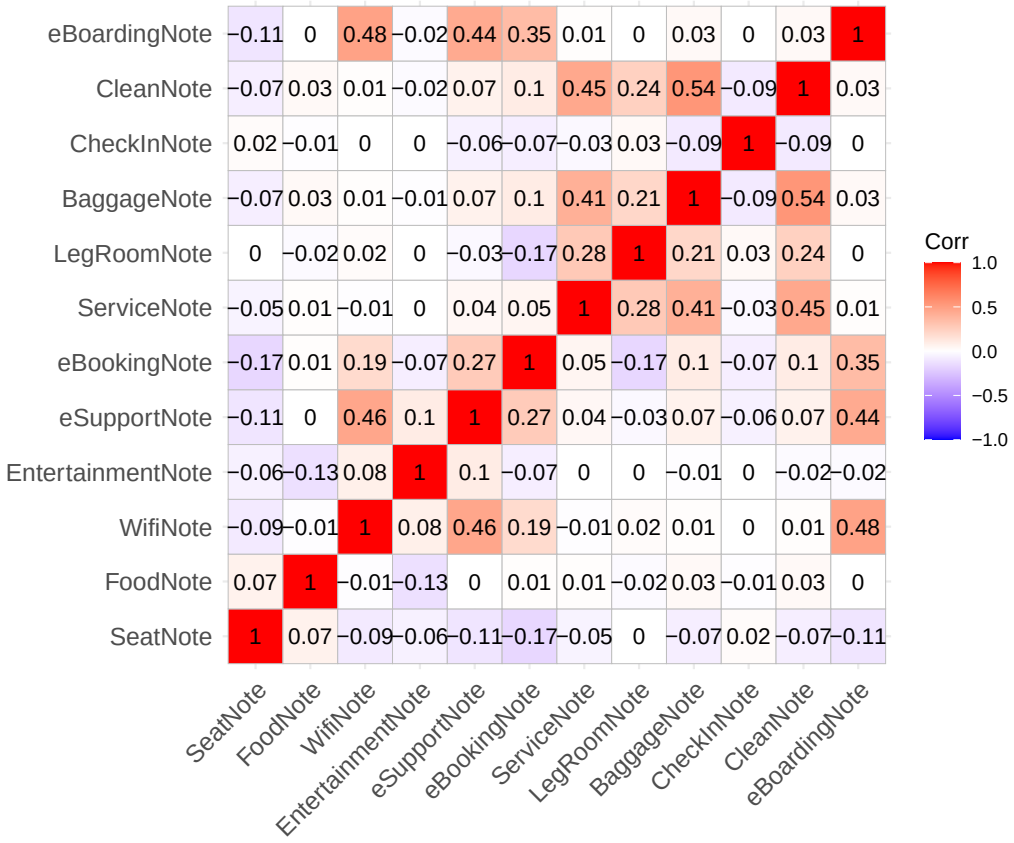


Figure 2: Spearman correlation shows no significant colinear relationships in ordinal variables

	Final.Bin	Total.Count	Total.Distr.	0.Rate	WOE	IV
1	<= 40	63722	51.6%	54.4%	-31.5	0.051
2	<= 59	48252	39.1%	34.4%	50.5	0.096
3	<= Inf	11580	9.4%	53.6%	-28.4	0.008
5	Total	123554	100.0%	46.5%	NA	0.155

WOE Table for DepartureDelay

	Final.Bin	Total.Count	Total.Distr.	0.Rate	WOE	IV
1	<= 19	99439	80.5%	44.7%	7.2	0.004
2	<= Inf	24115	19.5%	53.8%	-29.4	0.017
4	Total	123554	100.0%	46.5%	NA	0.021

2.2 Exploratory Data Analysis

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

3 The Logisic Regresssion

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.

#A train-test split of data

4 The performance of the Model

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 Validation of the Model

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 Monte Carlo Cross 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 The Challenger Models

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.

6.1 Neural Network

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.

6.2 Another logistic regression: logistic 2

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 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.

8 Bibliography