|  |
| --- |
| –– |
| “Life & Garden” - ADS-A 34 – Group B  Anjela Melkonyan, Ivan Germanov, |



Data Analysis Report

Anjela Melkonyan Ivan Germanov

Stijn Groenen Bart Verhagen

Ries van Geffen Redzhep Molaahmed

**Revisions**

|  |  |  |
| --- | --- | --- |
| **Version** | **Date** | **Changes** |
| 0.1 | 08-11-2018 | Initial start. Framework made. |
| 0.2 | 23-11-2018 | Written introduction, data, analysis, results and conclusions. Started on methods. |
| 0.3 | 24-11-2018 | Added churn EDA and Feature engineering section in Methods. |
| 0.4 | 25-11-2018 | Edited Introduction, Methods, added all EDA, Added prediction models, Conclusion, Recommendations and Appendix |
| 1.0 | 15-01-2019 | Divided sections – First and second analysis |
| 1.1 | 19-01-2019 | Documented findings, results and recommendations. Document finalized. |

**Distribution**

This document is sent to:

|  |  |  |  |
| --- | --- | --- | --- |
| **Version** | **Date sent** | **Name** | **Function** |
| 0.4 | 25-11-2018 | Rob Prop | Client |
| 0.4 | 25-11-2018 | Rafayel Ayetyan | Mentor - Machine learning |
| 0.4 | 25-11-2018 | Niek Schmitz | Mentor - Organizational context |
| 1.1 | 19-01-2019 | Rob Prop | Client |
| 1.1 | 19-01-2019 | Rafayel Ayetyan | Mentor - Machine learning |
| 1.1 | 19-01-2019 | Niek Schmitz | Mentor - Organizational context |

Table of Contents

[1. Introduction 3](#_Toc535703878)

[Reading guide 4](#_Toc535703879)

[2. Data 4](#_Toc535703880)

[3. Methods 6](#_Toc535703881)

[Data cleaning 6](#_Toc535703882)

[Feature engineering 7](#_Toc535703883)

[4. First analysis (2013 to 2017) 8](#_Toc535703884)

[Exploratory Data Analysis 8](#_Toc535703885)

[Churn analysis 9](#_Toc535703886)

[5. Models – first iteration 10](#_Toc535703887)

[Forecasting demand 10](#_Toc535703888)

[Churn prediction 12](#_Toc535703889)

[6. Second Analysis (2018) 13](#_Toc535703890)

[Exploratory Data Analysis 13](#_Toc535703891)

[Churn analysis 14](#_Toc535703892)

[7. Models – second iteration 15](#_Toc535703893)

[Forecasting demand 15](#_Toc535703894)

[Churn prediction 16](#_Toc535703895)

[8. Conclusions 16](#_Toc535703896)

[9. Recommendations 17](#_Toc535703897)

[10. Data ethics 18](#_Toc535703898)

[Introduction 18](#_Toc535703899)

[Related principles of Data Ethics 18](#_Toc535703900)

[Current situation 20](#_Toc535703901)

[11. Code solution and Jupyter Notebooks 21](#_Toc535703902)

[12. Appendix A: years 2013-2017 22](#_Toc535703903)

[13. Appendix B: Models - 1st iteration 29](#_Toc535703904)

[kNN 29](#_Toc535703905)

[Random Forest 30](#_Toc535703906)

[XGBoost 31](#_Toc535703907)

[14. Appendix C: year 2018 32](#_Toc535703908)

[15. Appendix D: Models – 2nd iteration 39](#_Toc535703909)

[kNN 39](#_Toc535703910)

[Random Forest 40](#_Toc535703911)

[XGBoost 41](#_Toc535703912)

# Introduction

The company Informa, a knowledge centre for the Netherlands and Belgium, has asked for help with empowering the management and officers of their client – who runs a business in the sector of Life & Garden – to make better decisions, direct actions based on trends, challenge the employees to adopt the best practices and identify and refine target audiences.

This study will show an analysis of all the data which the client has collected over the past few years. Results of this study will give conclusions, which can help the client to steer the organisation in a beneficial way. This document will outline the study, describe the methods used, present the statistics of the data, the different distributions of variables and give first conclusions from the developed predictive models.

With the investigation of the Exploratory Data Analysis (EDA), the team decided to focus on two machine learning models. These models will make valuable predictions which can to be useful for the business of Informa his client.

The first area to be further investigated is the forecasting of demand, the second one - understanding customer retention.

To solve those problems, we put in this exploration the current knowledge about the dataset at a glance. Questions that will be addressed are:

1. What demand of garden furniture can we predict for 2019?
2. What are the actual sales grouped by month?
3. What sales can we predict grouped by month?
4. What does the residuals per day look like?
5. When do we consider a customer churned?
6. Which features of the data will be the perfect predictors for understanding the customer churn?

With the results of EDA, the team noticed that the distribution of variables is mostly normal, and few realistic patterns can be seen. Therefore, the quality of the raw data collected must be improved.

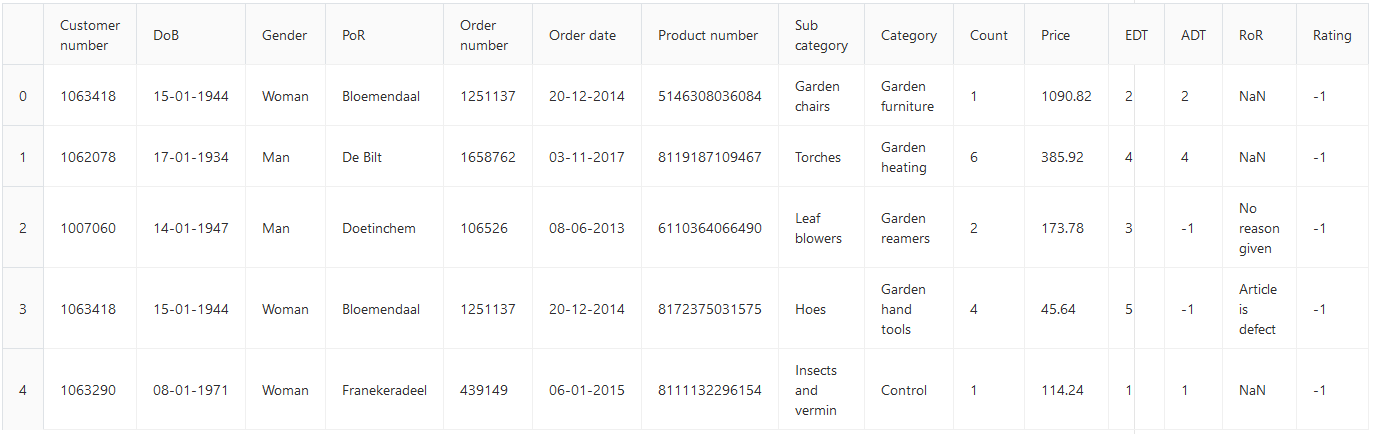
## Reading guide

The raw dataset is described in Chapter 2 of the document, including the assumptions the team made in the process to our predictions. The used methods are named and explained in Chapter 3. In Chapters 4 and 5 are the data analysis results and the models of the data from years 2013 to 2017. In Chapters 6 and 7 are the results and models for the year 2018.

Conclusion and Recommendations can be found in the end of the document, followed by 4 appendices with data visualizations. All Notebooks with code solutions will be delivered in additional files for further reference.

# Data

The dataset contains a list of orders from a web-shop in the field of Gardening and Lifestyle. It has 15 columns (Figure 1) and was given in two sets. The first one was with data for 5 years (from 2013 to 2017) and the second one was for around 9 months (in 2018).



**Figure 1** First 5 records in the dataset

The type of every column is read from the raw dataset as follows:

* Customer number: object,
* DoB (Date of birth): object
* Gender: 'category'
* PoR (Place of residence): object
* Order number: object
* Order date: object
* Product number: object
* Sub-category: 'category'
* Category: 'category'
* Count: np.int64
* Price: np.float64
* EDT (Estimated date of arrival): np.int64
* ADT (Actual date of arrival): np.int64
* RoR (Reason of return): object
* Rating: np.int64

To begin the operation of the project, the team has assumed that the company operates in the Netherlands and Belgium.

# Methods

A general overview of the activities performed on the data is:

* Data cleaning
* Feature engineering
* Exploratory Data Analysis
* Creating Forecasting demand model
* Analysing Churn and related features
* Creating Predicting churn model

Therefore, Python and related libraries (Pandas, NumPy, Seaborn, Matplotlib, sklearn, etc.) were used.

## Data cleaning

To determine if the data has to be cleaned and what exactly has to be cleaned, a few functions were applied:

* Counting the null values in the dataset (in every column)
* Counting and displaying the unique values in a column to determine if there are white spaces or new lines, mistakenly written in the dataset
* Transforming date columns from object to datetime, so functions can be applied to them further in the feature engineering
* Transforming all dates to one format (d-m-Y)

After applying these algorithms, the problems found were the following:

* Only one column contains null values – RoR (Reason of return). This is understandable as the dataset represents a list of order lines and neither all orders/product were returned nor for all of them the customer has written a reason for return. However, the number of these null values is a lot - 4,303,432 out of 4,523,276. The solution for this missing data is introduced in the Feature engineering section.
* There are also missing values in the Gender column, but they are considered statistically unimportant and a solution to this will also be introduced in the Feature engineering section.

## Feature engineering

**Feature engineering** is the process of using domain knowledge of the data to create **features** that make machine learning algorithms work. The team engineered several features from the dataset, which were useful for the predictions.

With regards to the forecasting demand issue, the following features were engineered:

* Man/Woman – by one-hot encoding the Gender column
* Price per product
* Age – from date-of-birth of customers

The reason behind using One-hot encoding for the Gender column is that when a direct substitution of the gender categories with 0s and 1s is done, some algorithms consider the 1 more than the 0. This means that there will be gender preference done by the algorithm. Therefore, the One-hot encoding separates the Gender categories in two separate columns and gives them the proper value of 0 or 1.

The price per product is calculated by dividing the total price of an order line to the quantity of the product ordered. The age is calculated by subtracting the DoB (Date of Birth) from the date of the analysis (23/11/2018).

When it comes to the model predicting churn, the following features were added:

* Orders per customer
* Average rating per customer – does not consider orders where no ratings were given
* Average time between orders
* Estimated time of arrival/Actual time of arrival – proportion, takes value of 0-1
* Percentage returned items per customer
* Churn – categorical variable, taking a 0 or 1 value

Churn is the most important featured that was engineered, because it is also the feature that is to be predicted. Defining customer churn for a business strongly depends on how the strategic management wants to define it. A good solution to define if a customer has been lost is to see if he/she has not bought a product from the web-shop for a period of time. That period of time depends on the customer. A customer is defined as having churned if he/she has not ordered anything for a longer time than the avg. time between his/her orders + 1 std. of his order days. Example customer A has ordered 5 times in total, meaning we have the days between his 1st order and 2nd order, his 2nd order and 3rd order, 3rd order and 4th order, 4th order and 5th order, 4 days in total. When we calculate customer A’s avg. time between orders, it is 100. When we calculate the standard deviation of those days, it is 50. That means if he/she has not ordered anything for longer than 100 + 50 = 150 days, he is defined as having churned. If he has ordered, he is defined as not churned.

# First analysis (2013 to 2017)

## Exploratory Data Analysis

#### Data

The first data provided by Informa is a collection of records and information about customers and their orders for the years from 2013 to 2017.

#### Results

The analysis on the first set of data we received did not show any specific evidence of patterns or target groups that the company and its marketing department should focus on.

According to the gender distribution in the dataset, the orders are mainly made by male customers with 69.9%. (Figure 5, [Appendix A](#_Appendix_A:_years)) Analysing the ages of the customer, shows that there is a big peak in the ages between 40 and 55, which suggest that this is the main target group of the company. (Figure 6, [Appendix A](#_Appendix_A:_years))

If we move on to the orders’ details, we can see a total of 12 categories of products, where one category stands out – ‘Gardening furniture’ with a revenue of €1,065,657,468.92. More information about the Categories and subcategories can be found in the Notebook solutions provided with the deliverables, but considered they are not helpful for the analysis, they will not be included in this report.

Calculating the total number of orders per customer shows that there is no outstanding information that can be collected from the data. (Figure 8, [Appendix A](#_Appendix_A:_years))

In addition, with the analysis performed, we tried to find a correlation between the price of a product and the quantity a person orders at once, but such was not found, as well as correlation between price of a product and rating given. (Figure 9 and Figure 10, [Appendix A](#_Appendix_A:_years))

A part of the analysis worth mentioning is the comparison between Actual date of arrival and Expected date of arrival per order, which shows there is a difference of around 1 day between the two days. Therefore, we can conclude that there is no problem with the logistics department.

Finally, the percentage of returns and the average rating per customer were analysed and showed that customers return their ordered items maximum 15% of the time. In addition, the customers’ ratings are normally distributed, but a small peak can be seen around the rating value of 4.

## Churn analysis

To plot every variable in comparison with the churn and fully analyze the features’ relation to the churn values, a new dataset was created combining the following features (per customer):

* Customer number
* T\_between\_orders
* Number of orders
* Average raiting
* Percentage returns
* EDT
* ADT
* EDT\_divided\_ADT
* Man
* Woman
* Age
* Churn

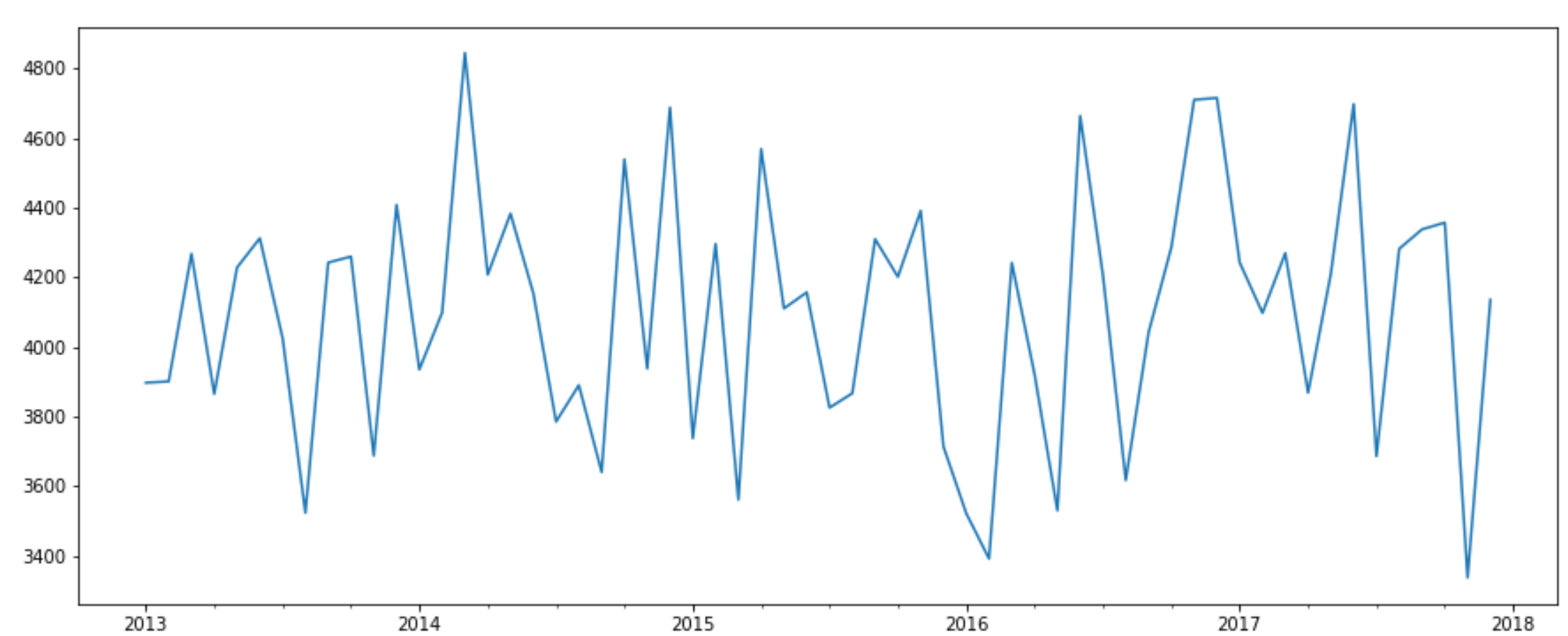
As the previous chapter already discussed how the churn formula was created, we discovered that on the first 5 years of the data provided, only 17.1% of the customer churned, which is about 15k customers. (Figure 12, [Appendix A](#_Appendix_A:_years)) As 17.1% might not seem much, when speaking about churn, every percentage is important. Therefore, the reason of this analysis is to determine which features can be good predictors for a customer churn model and how customer churn might be prevented.

The following variable analysis looked into the distribution of the numerical variable values in the customer churn feature. Is there a feature highly correlated to customer churn? The answer found is ‘No’. Looking into the relationships between the time between orders, number of orders, average rating per customer, percentage of returns and the churn value, we can conclude that there is no such factor, which can separate customers that churned and the ones that were retained. (Figure 13, 14, 15 and 16, [Appendix A](#_Appendix_A:_years)) To prove that, a correlation matrix was also visualized. The only small correlation noticed was between churn and the time between orders, which is obvious as this time was used in the formula for churn.

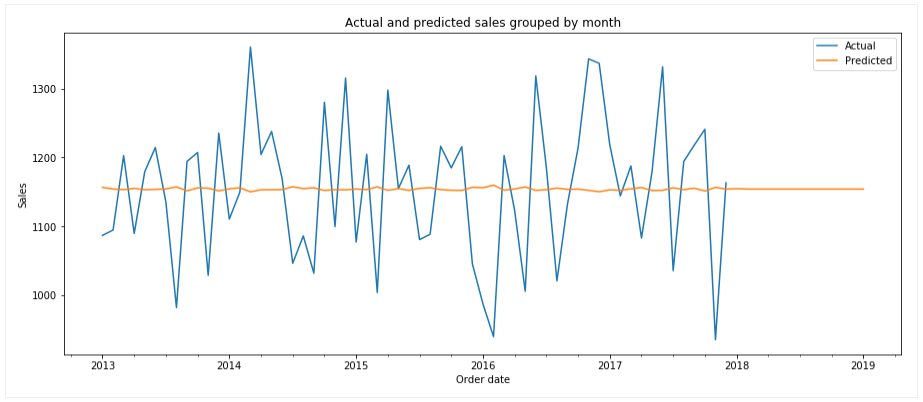
# Models – first iteration

## Forecasting demand

Figure titled *Orders by season* shows that there are not patterns in the sales. This makes it very hard to create an accurate model. The results in figure titled *Orders by category* show that the predictions are around the mean.



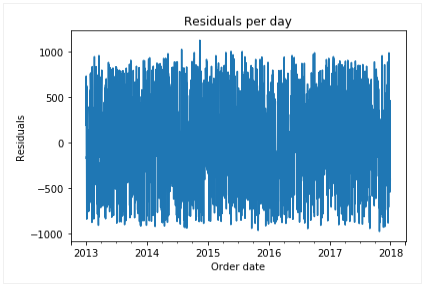
**Figure 18** Orders by season



**Figure 19** Category by season

#### Residual errors

The plot of *Residuals ordered by date* gives a better look in the differences between the actual sales and the predictions.



**Figure 20** Residuals ordered by date

The applied ARIMA model has an error rate of 291042.419 which is useful to know to compare to other models. The error rate should be as low as possible.

In the last two figures is shown that the error rate mostly is within one thousand sales, which is not really good.

## Churn prediction

For the churn prediction, a new dataset was built, as mentioned in the Analysis chapter. The futures it consists of are:

* Customer number
* T\_between\_orders
* Number of orders
* Average raiting
* Percentage returns
* EDT
* ADT
* EDT\_divided\_ADT
* Man
* Woman
* Age
* Churn

As shown in the analysis, only 17.1% of the customers has churned, based on the method for calculating churn.

As the data is not relevant enough for predictions, the best out of 6 prediction models were trained – kNN, Random forest and XGBoost. In addition, as the churn feature is imbalanced, we used the SMOTE algorithm. At a high level, SMOTE creates synthetic observations of the minority class (not churned customers) by:

1. Finding the k-nearest-neighbours for minority class observations (finding similar observations)
2. Randomly choosing one of the k-nearest-neighbours and using it to create a similar, but randomly tweaked, new observation.

The results of the model can be seen in [Appendix B](#_Appendix_B). To explain them in short, the accuracy of the models was the following:

* [kNN](#_kNN) – 61%
* [Random Forest](#_Random_Forest) – 80%
* [XGBoost](#_XGBoost) – 81%

The models cannot be judged by the accuracy score, because the data was imbalanced and on the model performance reports in Appendix B. Thus, the precision and recall score help us determine the performance of the models and it can be seen that they are also with a low value.

# Second Analysis (2018)

Exploratory Data Analysis

**Data**

The second set of data provided by Informa is a collection of records and information about customers and their orders for 9 months in 2018.

**Results**

The analysis on this data we received shows a bit more interesting result.

The gender distribution is absolutely the same as in the first set of data (Figure 5, [Appendix C](#_Appendix_C:_year)), as well as the ages of the customers (Figure, [Appendix C](#_Appendix_C:_year)).

The category of ‘Gardening furniture’ is with the highest revenue again of €152,726,203.84 only for nine months (Figure 7, [Appendix C](#_Appendix_C:_year)). After a closer analysis it was discovered that there is around 30% less revenue comparing to the average of 2013 to 2017.

Another part of the analysis worth mentioning is again the comparison between Actual date of arrival and Expected date of arrival per order. In the last year, the data shows that the expected delivery time has gone up with a day and a half, but the actual delivery time has gone up with 3 days (Figure 11, [Appendix C](#_Appendix_C:_year)). That shows a big difference between the estimated and the actual time of delivery. This fact suggests a problem occurring at the logistics department during year 2017.

The analysis of the returns, ratings and time between orders did not show any valuable results.

Churn analysis

The churn feature in the second set of data given is much more valuable for the models. In addition, the numbers look interesting on the first sight, but only without considering that the time frame is only 9 months. The customers considered churned in the last 9 months is 49.8% - almost half of all the customers in the dataset (Figure 12, [Appendix C](#_Appendix_C:_year)).

The analysis of the distribution of numerical features in the churn variables did not show anything that can help separating churned from non-churned customers. Nevertheless, call visualizations can be seen in [Appendix C](#_Appendix_C:_year) (Figure 13, 14, 15, 15 and 17). They are a proof that the features analysed are with the same values, both in churned and retained customers and cannot be beneficial for the models created later on.

# Models – second iteration

## Forecasting demand

Forecasting can be useful for insight in what people will mostly be ordering in the future.

The model developed forecasts by day. In the beginning, the model was meant for forecasting sales demand in general, but later on it was decided to follow one category at a time. The results of the graphs, both of the ways, were similar. The reason behind this might be that all the sales are evenly spread over the different categories.

A close up of a logo

Description generated with very high confidenceFigure 21 shows the sales made in the garden furniture category. The blue line represents the actual sales based on our data. The orange line represents the predictions from our ARIMA model.

**Figure 21** Orders by season

We combined the datasets into one sales dataset. We took all sales in the garden furniture category and then we created a new dataset with sales per day. This is the dataset we used to train our ARIMA model. We added all sales together for each month to create this graph. We did the same for the predicted sales and added the results to the graph.

As you can see the ARIMA model came up with a pattern based on the actual sales in the past. The troughs in our predicted pattern are close to troughs in the actual sales data. Unfortunately, the rest of our predicted pattern seems quite random. We think that the problem is that the actual sales data has no clear patterns in it for our model to train on.

## Churn prediction

The second iteration of the churn prediction models’ development, the algorithms were trained on a merged set of data – the whole 6 years, as running the models through only 9 months of data won’t show any good results.

The 3 algorithms mentioned in the previous iteration has scored the following accuracy:

* [kNN](#_kNN_1) – 55%
* [Random Forest](#_Random_Forest_1) – 68%
* [XGBoost](#_XGBoost_1) – 72%

As mentioned before, the accuracy score will not be only value to determine the performance of the models. With a certain amount of parameter tuning, the algorithm scored higher recall and precision scores. All model performance reports can be seen in [Appendix D](#_Appendix_D).

# Conclusions

After two iterations of data analysis, there were no clear benefits of the data:

* There is no correlation between quantity and price
* Almost all ages from customers contain the same number of customers
* There is no correlation between rating and price
* There is no correlation between orders and season
* There is no correlation between categories and season
* The correlation matrix shows that there are no variables with high correlation that can be used as good predictors for the models

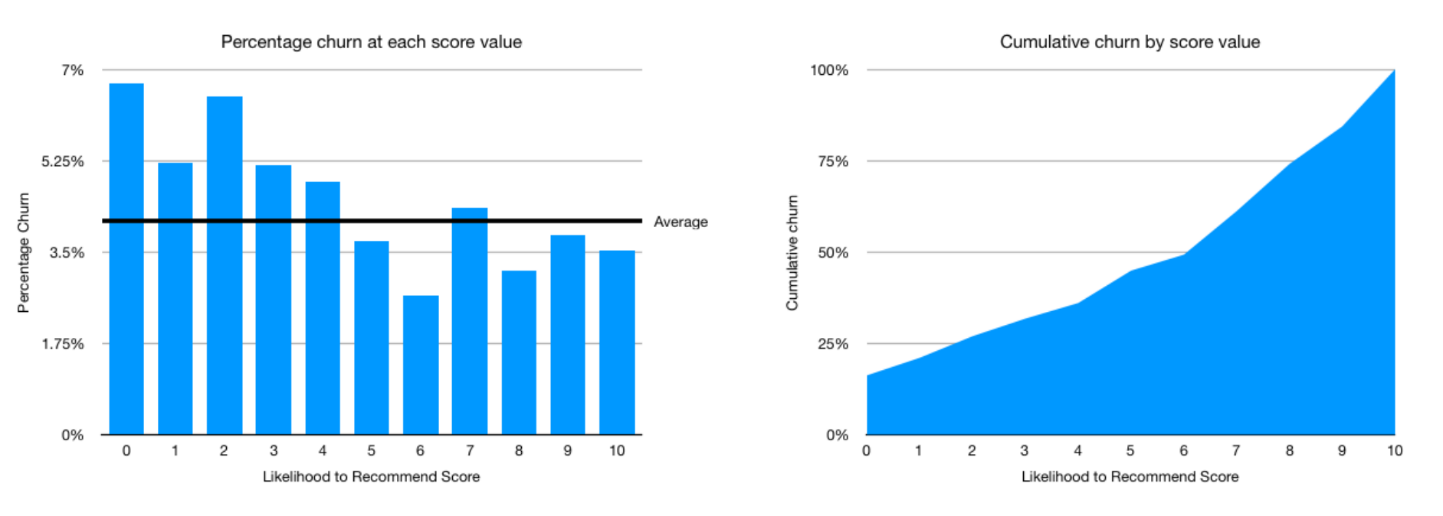
If there were patterns to find in the sales numbers from the given dataset, then the predictions would be more accurate. When evaluating the Forecasting demand model, it can be concluded that the model predicts sales around the total sales mean for every category. This points to the fact that seasons did not have any effect on which categories were sold more or less.

In regard to the churn model predictions, it looks like the attempted models are accurate and can be beneficial for the business goals of the client. Although the steps taken to develop the models led to getting the precision and recall scores to their maximum, deeper analysis shows that these scores are proof of the bad performance of the model. Of course, collecting more data in the years will lead to better performance of the model.

# Recommendations

Even though some of our predictions were unsuccessful because of the data’s limitations, there were still useful patterns that we found, which could be useful for the client. For instance, the distributions of orders per customer, percentage of returns per customer, average days between the orders per customer, etc. A valuable finding is the huge difference between the expected date of delivery and the actual one. This suggests problems in the logistics department, which have to be addressed as soon as possible. However, much more business and especially marketing value can be extracted from the data analysis visualizations performed.

In terms of data gathering, we recommend that the client collects more data and especially data in which patterns can be found. Otherwise, building models for predictions will yet again prove to be a difficult task. New data sources can be used on top of the usual transactional CRM data (profile and purchase history) to improve the performance of the predictive models - service usage, client feedback and customer service requests, social media interactions, web tracking data...can all be revelatory of customer churn.

In addition to this, customer feedback can be collected not only to discover problems in the services but combined with prediction models, that can be a really powerful tool. To demonstrate an example, we made the following graph:

Analysis like this one connects the scores of a feedback session with churn prediction. This visualization shows how untrustworthy the feedback scores can be. As seen, the percentage of customers, who gave a score between 0 and 4 is much higher than the rest. However, even the percentage of people who gave 10 is close to the average (almost 4% of the customers). For a customer churn value, this is an amount to consider!

Customers want to feel valuable. A personalized response after a customer has given feedback might also be a good idea. An internet research shows that one of the most correlated features with customer churn is the loyalty mentions of the clients.

# Data ethics

## Introduction

As Data Science provides big opportunities for the improvement of people’s public and private life, as well as the environment around us, unfortunately these opportunities create additional ethical challenges. The extensive use of data, also personal in this case, the automation of all processes and the reduction of human involvement or oversight, poses pressing issues of responsibility, ownership, privacy, human rights, etc.

Therefore, in this chapter will address each one of the principles of Data Ethics related to the project and take them into consideration.

## Related principles of Data Ethics

#### Ownership

Individuals own their data. Personal data refers to data sets describing a person ranging from physical attributes to their preferences and behaviour. In this project, data taken into consideration includes: date of birth, place of residence, products ordered, date of orders, written reason of return and rating.

#### Transaction Transparency

Customers should have a transparent view of how, why and for what their data is being used. The individual has the right to know:

1. Why the data is being collected? – It is being collected for business analysis and overview, making possible for the organization to improve its commercial actions, management and logistics services in order to serve the customer better.
2. How is it going to be used? – This data is used for the analysis and creating of prediction models that will help the organization and its management take decisions in making the service better for the customer.
3. How it can be amended by the individual concerned? –The data that is not absolutely necessary for the company’s services to be achievable can be updated or deleted upon request from the individual.

#### Consent

User’s data is utilized only after the customer gives consent to their data being used.

If an individual or legal entity would like to use personal data, the organization should inform the owner of the data and explicitly express what personal data moves to whom, when, and for what purpose. In addition, the owner of the data should have the ability to manage the flow of their private information across massive, third-party analytical systems.

#### Privacy

Private customer data and identity should remain private. This principle does not imply secrecy. However, data obtained from a person should not be shared and exposed to any other businesses or individuals with any trace to person’s identity.

Security

Security can be used to combine the previously discussed principles. It engages with:

**confidentiality** – the organization must assure the security of data so that it is only available for authorized party

**availability** – the organization should ensure in reliable manner only for authorized users to access personal data, the Website as well as any E-commerce data service

**integrity** – the organization should make sure that data is collected accurately and truthfully

**authentication** – the organization should confirm that the person or other parties who are using the data is the one who they claim to be

**nonrepudiation** – the organization should guarantee that online customers or commerce partners cannot be wrongly denied accessing to the data

## Data Ethics - Current situation

Concerning Data Ethics, this project answers the following questions:

1. Why is my data being collected?

* Improving logistics
* Improving demand and supply,
* In general, improving business processes

1. How is my data going to be used?

* Prediction models
* Descriptive Analytics

1. How it can be amended by the individual concerned?

* The ability of the customer customer to not give or delete their data

1. User’s data is utilized only after the customer gives consent to their data being used.
2. Data obtained from a person should not be shared and exposed to any other businesses or individuals with any trace to person’s identity

# Code solution and Jupyter Notebooks

The Python notebooks for the following actions will be delivered:

* Preparing and Cleaning
* Merging the second set of data to the first one
* Analysis on the first set of data
* Analysis on the second set of data
* Model of forecasting demand
* Model of churn prediction

They can be referenced for further technical questions.

# Appendix A: years 2013-2017

**A screenshot of a cell phone

Description automatically generated**A picture containing compact disk, electronics

Description automatically generated **Figure 5** Gender distribution in orders

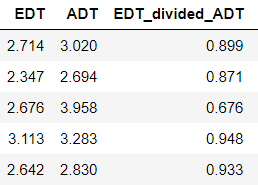
**Figure 6** Age distribution in orders

A close up of text on a white background

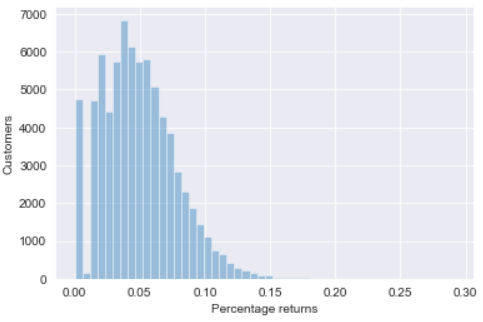
Description automatically generated**Figure 7** Money earned per category

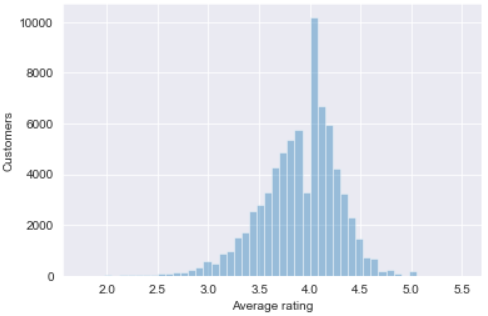
A screenshot of a social media post

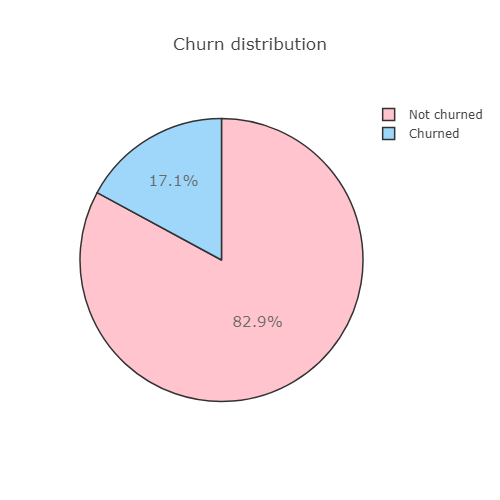
Description automatically generated**Figure 8** Distribution of number of orders per customer

A screenshot of a cell phone

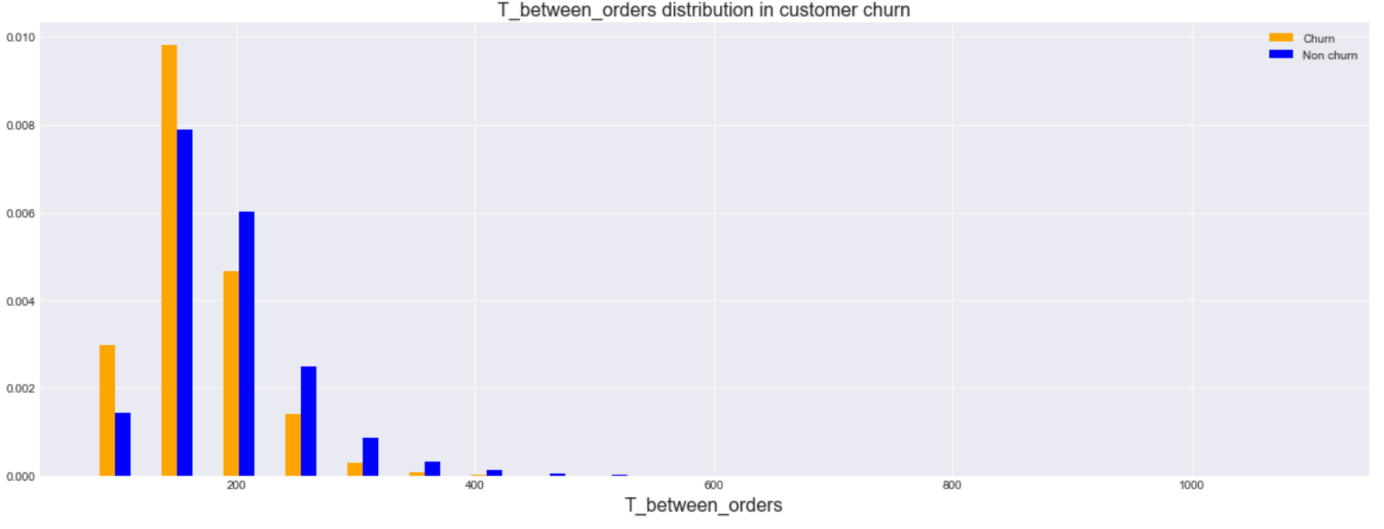
Description automatically generated**Figure 11** Avg. EDT vs Avg. ADT

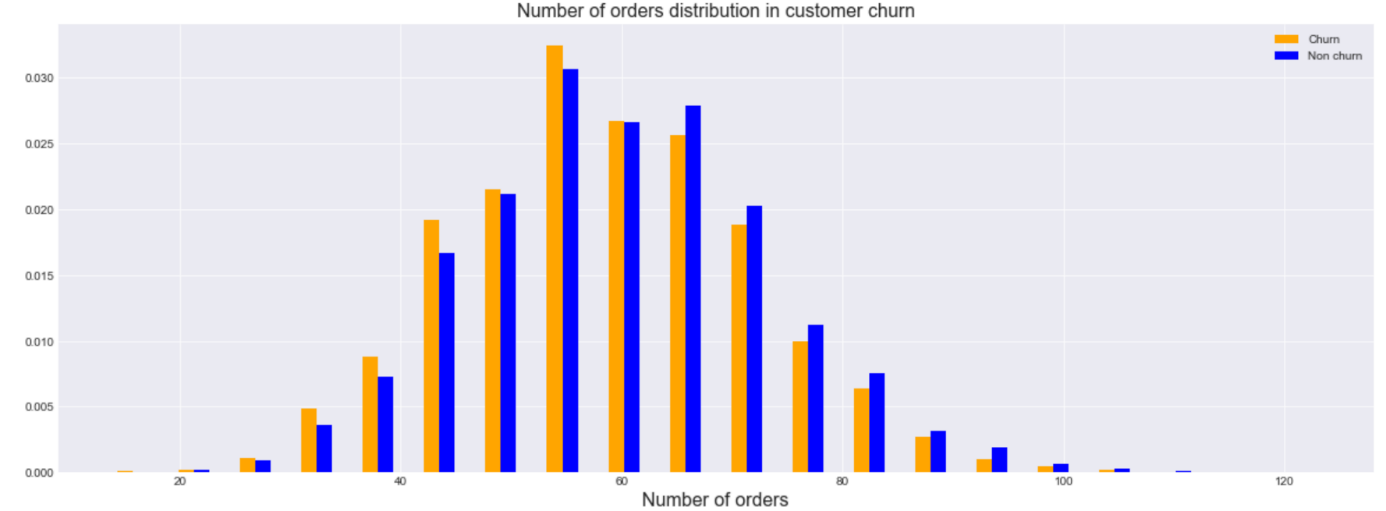
**Figure 14** Percentage returns/customer

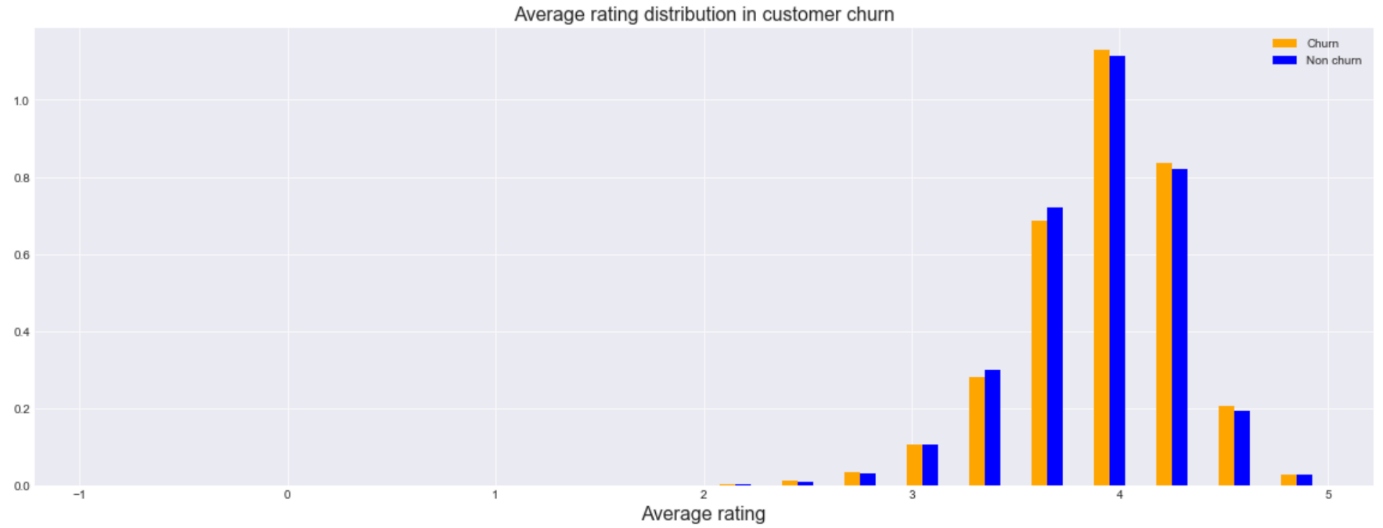
**Figure 2** Average rating/customer

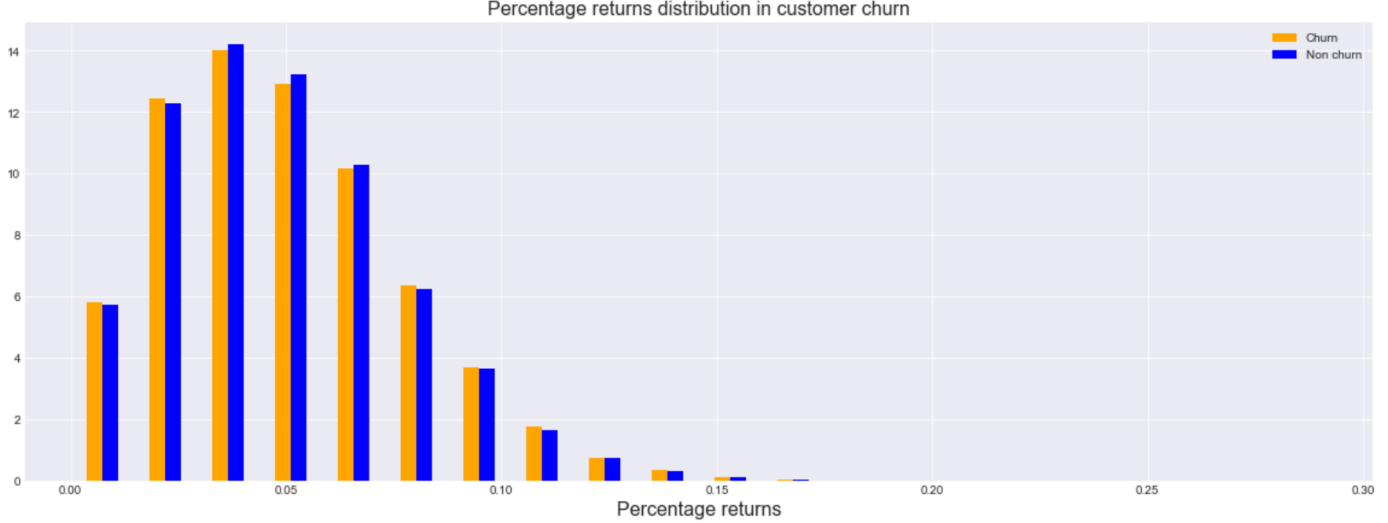
A screenshot of a video game

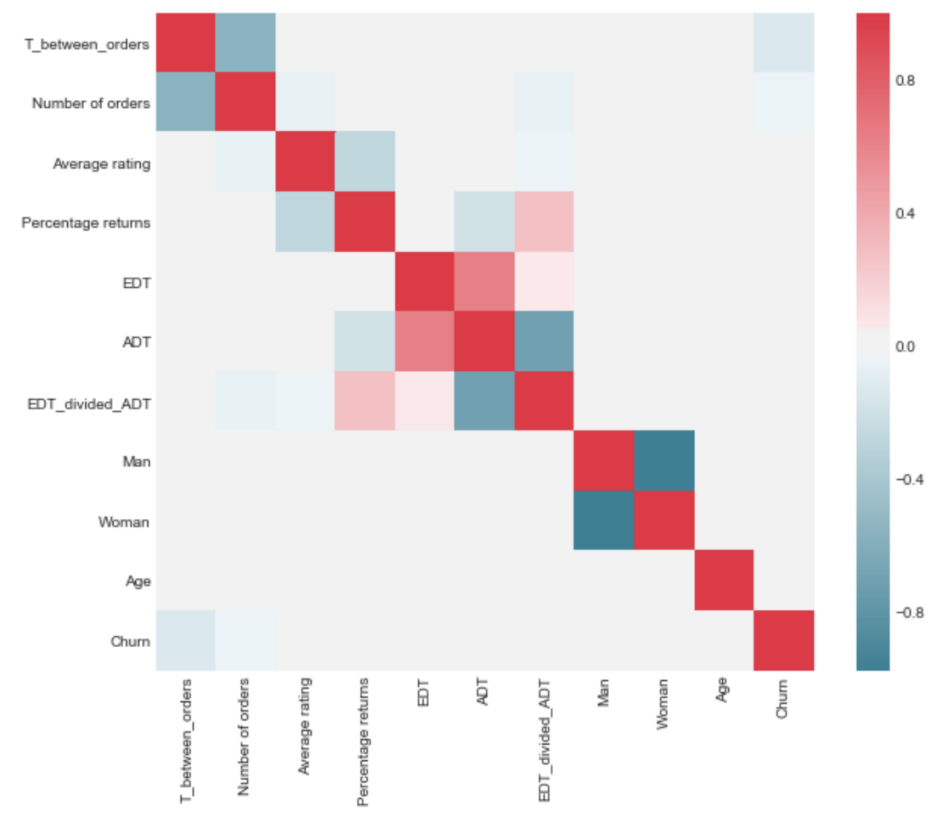
Description automatically generated**Figure 12** Proportion Churned/Non-churned customers

**Figure 13** Time between orders in relation with his/hers churn value

**Figure 14** Number of orders per customer in relation with his/hers churn value

**Figure 15** Average rating per customer in relation with his/hers churn value

**Figure 16** Percentage of returns per customer in relation with his/hers churn value

**Figure 17** Correlation matrix of introduced features

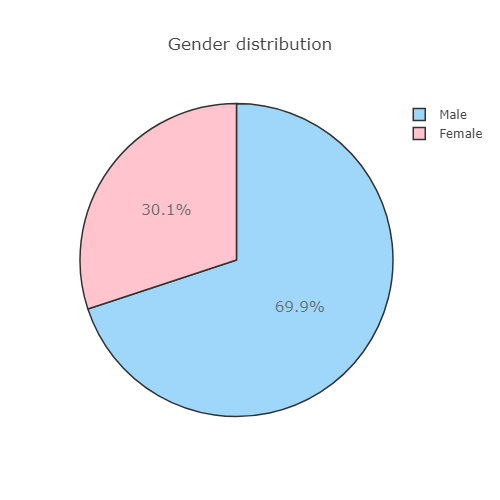
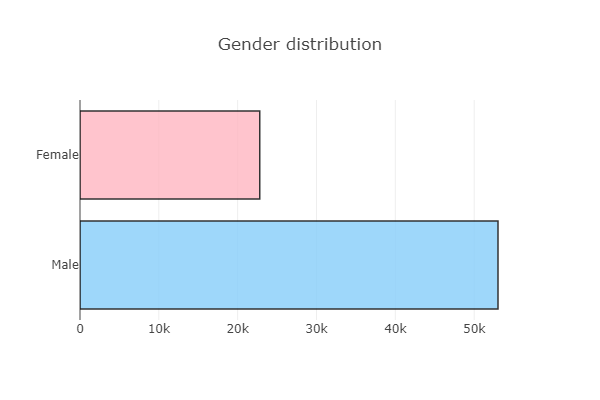
# Appendix B: Models - 1st iteration

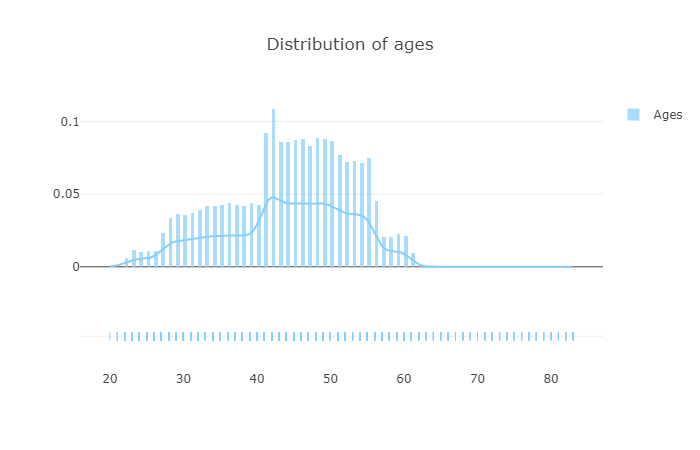
## A close up of a map Description automatically generatedkNN

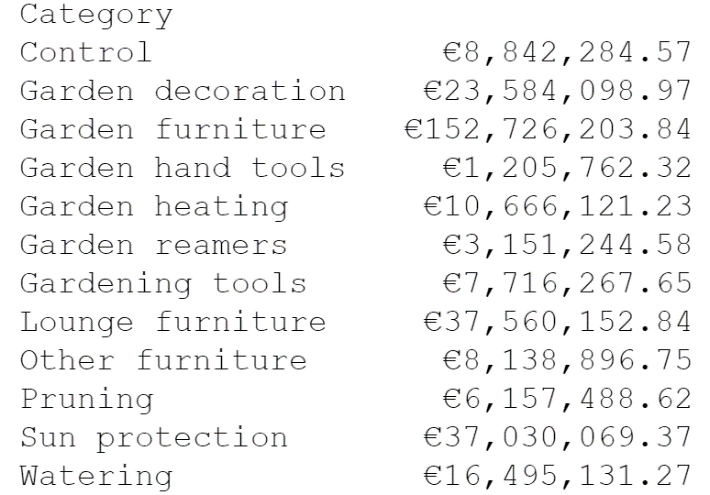
## Random Forest

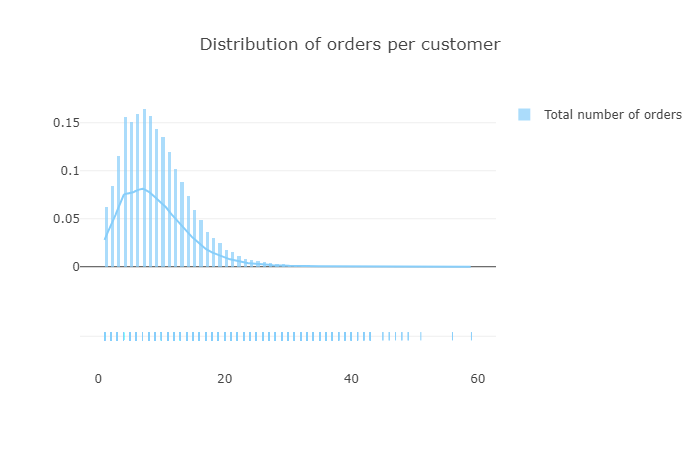
## XGBoost

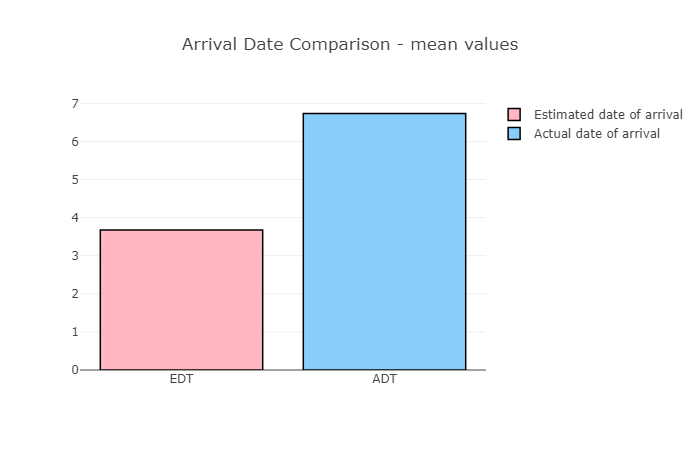
# Appendix C: year 2018

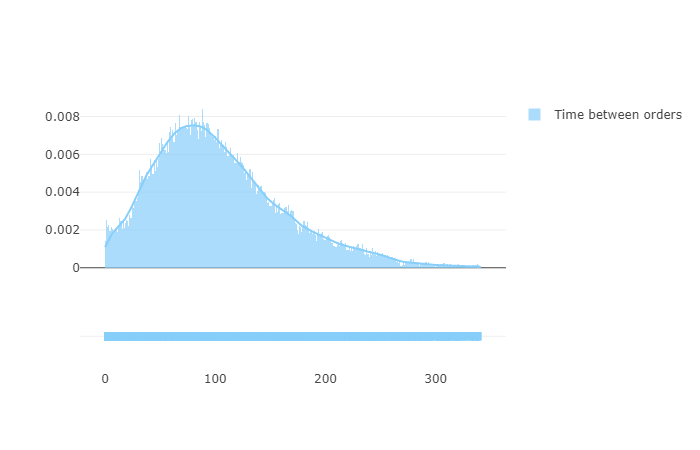
 **Figure 5** Gender distribution in orders

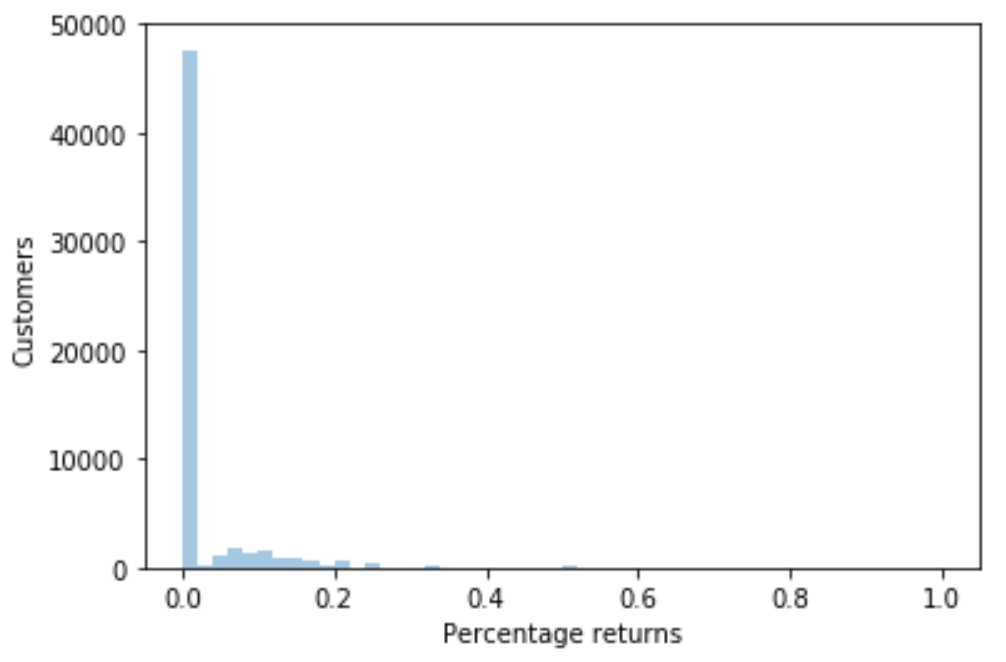
**Figure 6** Age distribution in orders

**Figure 7** Money earned per category

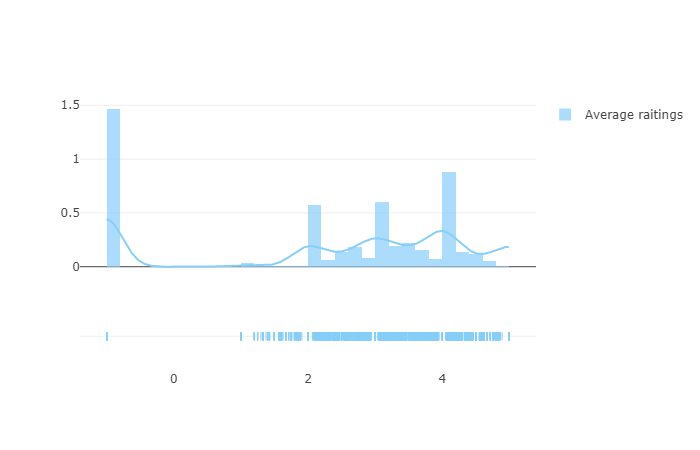
**Figure 8** Distribution of number of orders per customer

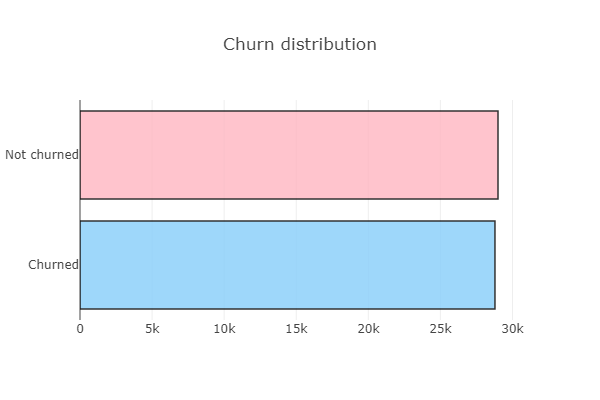
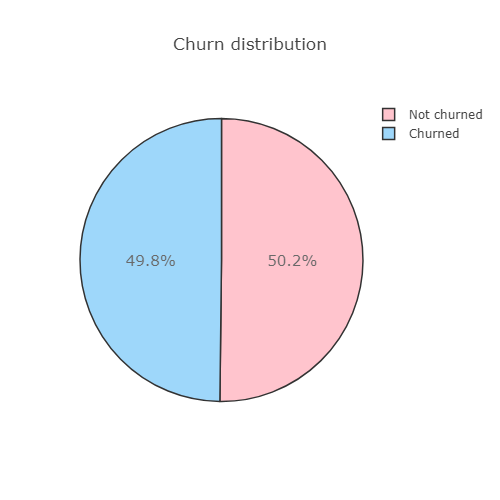
**Figure 11** Avg. EDT vs Avg. ADT

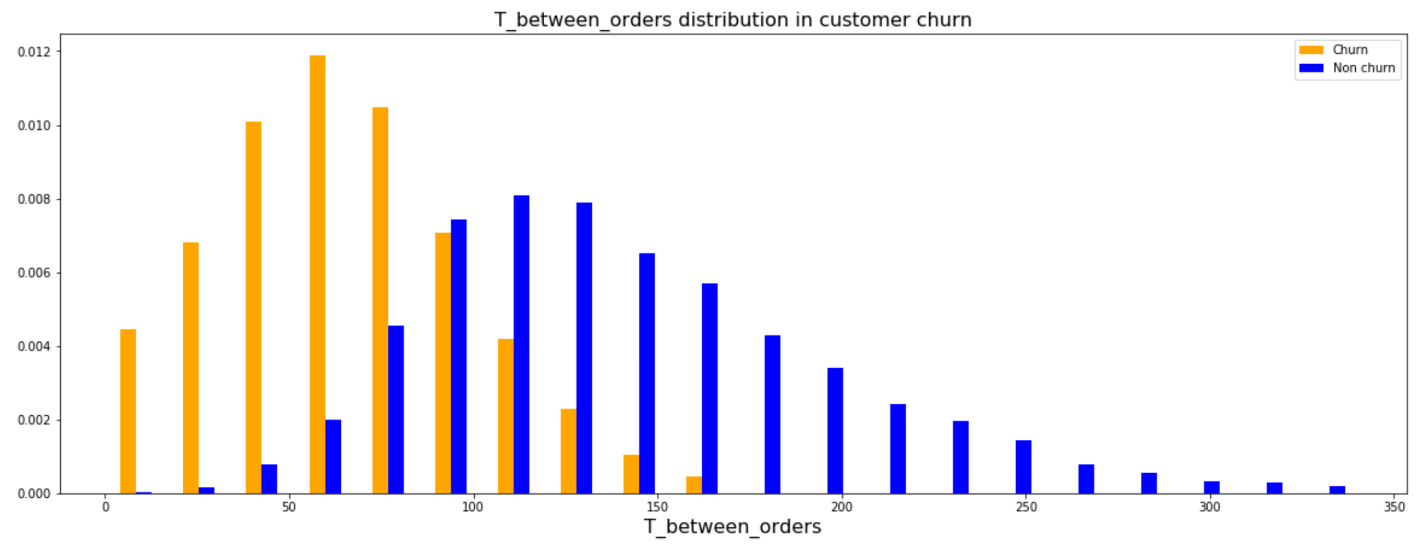
**Figure 33** Average time between orders for customers

**Figure 14** Percentage returns/customer

**Figure 4** Average rating/customer

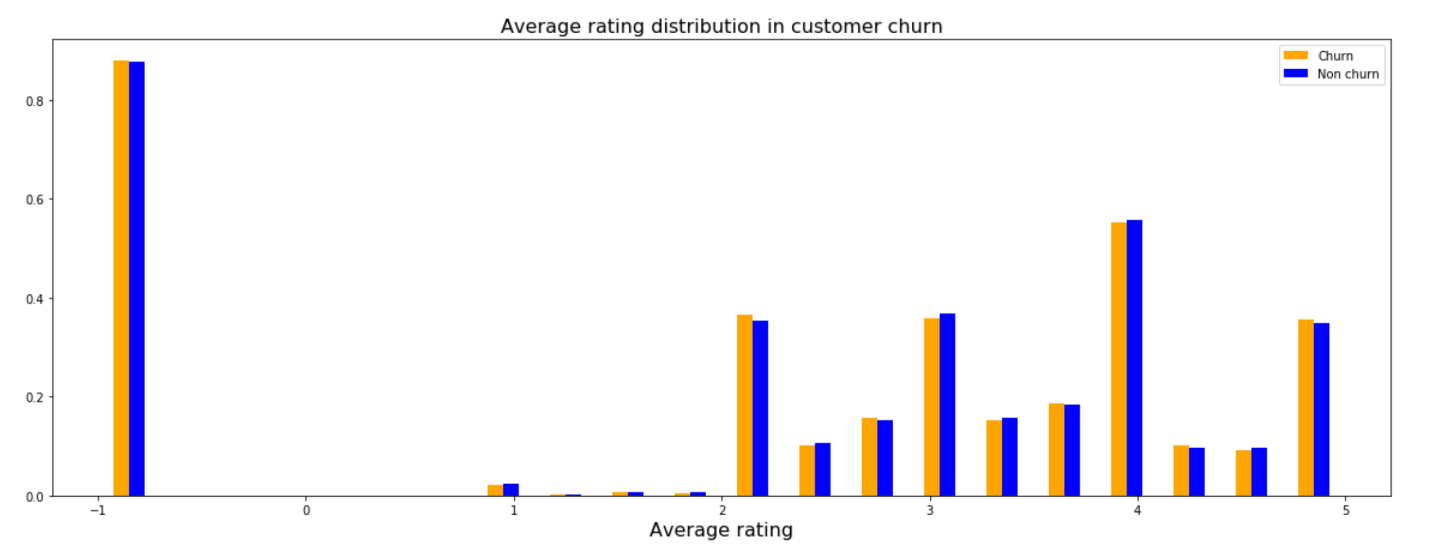
****

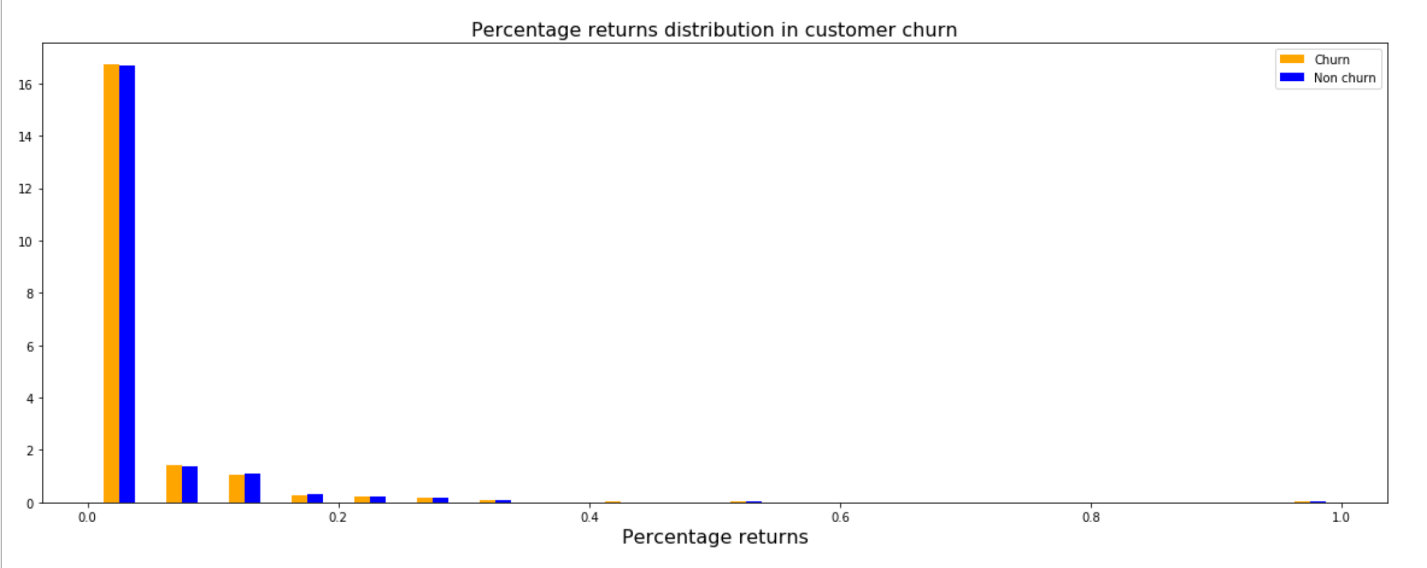
**Figure 12** Proportion Churned/Non-churned customers

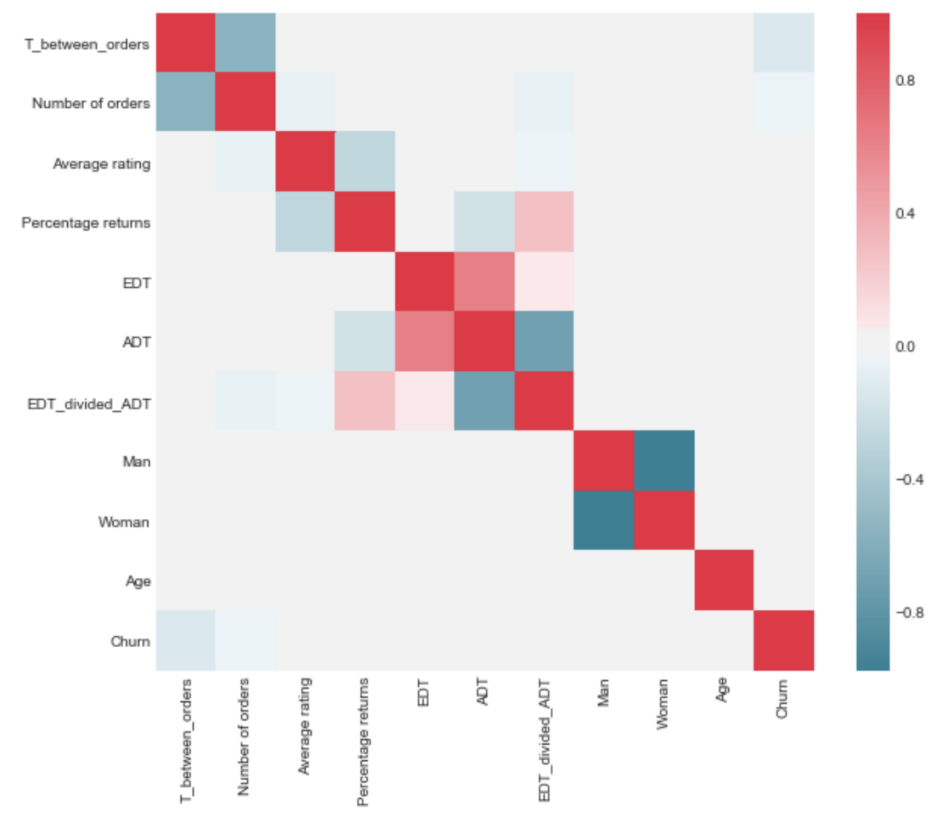
**Figure 13** Time between orders in relation with his/hers churn value



**Figure 14** Number of orders per customer in relation with his/hers churn value

**Figure 15** Average rating per customer in relation with his/hers churn value

**Figure 16** Percentage of returns per customer in relation with his/hers churn value

**Figure 17** Correlation matrix of introduced features

# Appendix D: Models – 2nd iteration

## A close up of a map Description automatically generatedkNN

## Random Forest

## XGBoost

