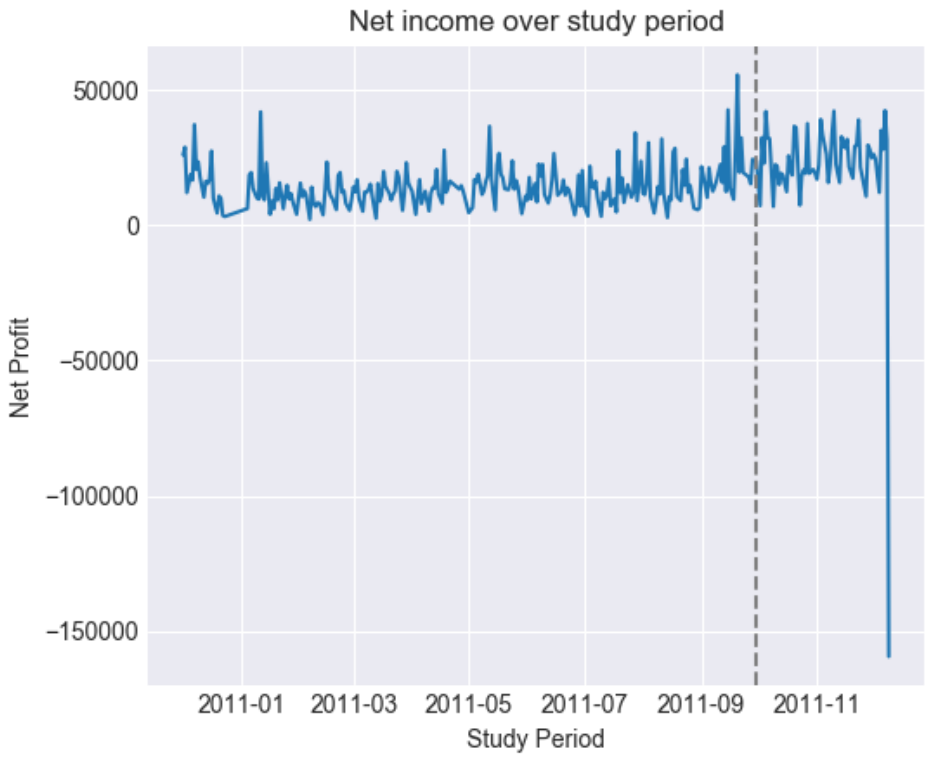
Submission for HelloFresh Internship in Data Science
Applicant Full Name: Ilias Katsabalos
Email: i.katsabalos@gmail.com
Mobile: +306947538202

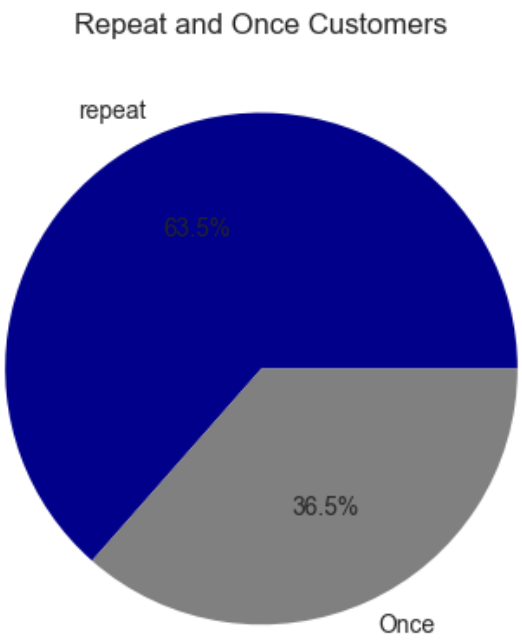
1. Exploratory Data Analysis

I used python for the exploratory data analysis and imported the pandas, numpy, matplotlib and seaborn libraries, all included in the latest version of anaconda. The file attached is called "Ilias_Katsabalos_*EDA.**ipynb***" and can be opened with the jupyter notebook. I also included "Ilias_Katsabalos_*EDA.**py***" but I suggest you run the **ipynb** version, as it is much more self-explanatory. For any occasion, I will present all the visuals here also, for a quick look.

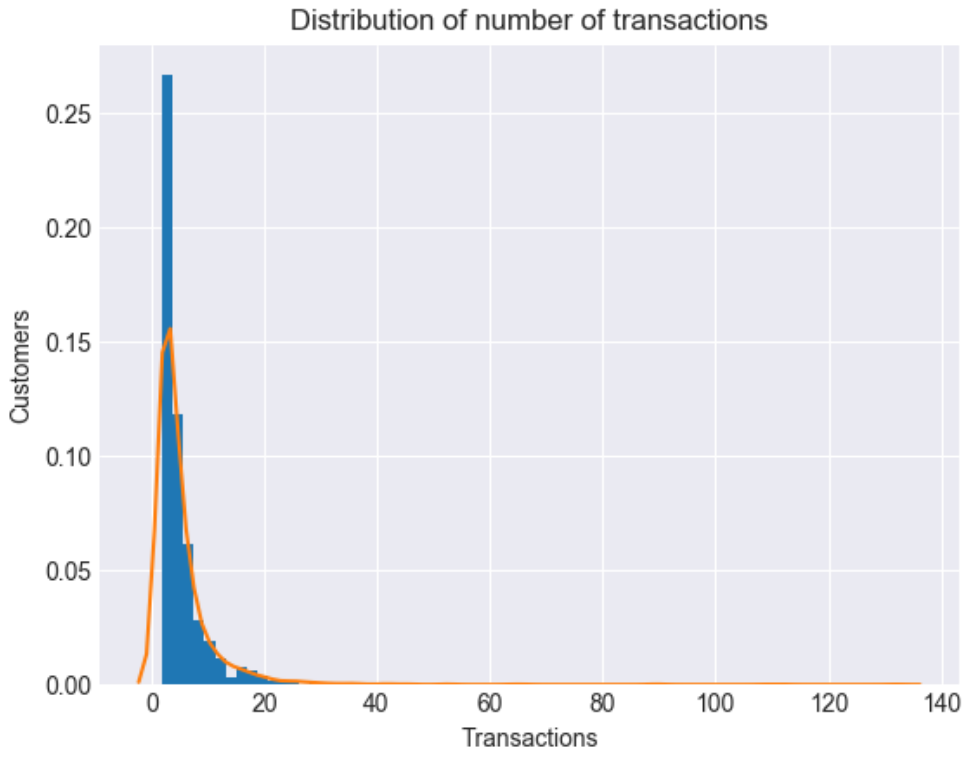
The first thing to do, is to observe the purchases, over the period of time and check for a good point to split the data in train and test sets. In the diagram below, there is not a significant difference between the train and test period (The split is indicated by the dotted gray line). The big outlier is studied more carefully in the python script.



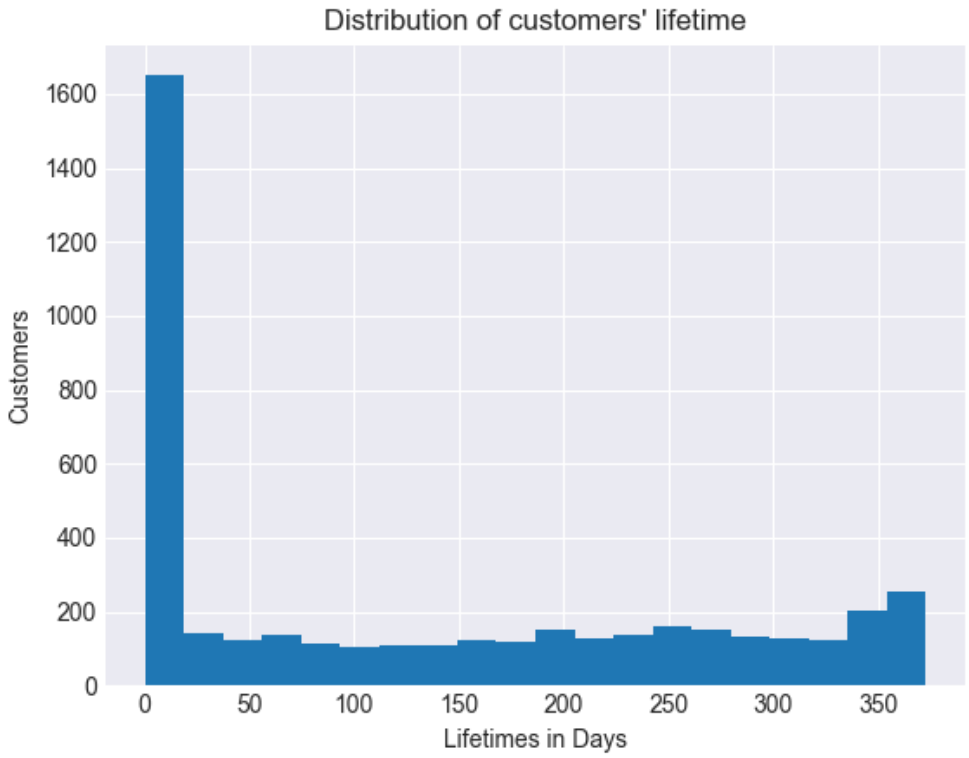
What I did next, was to find the percent of customers which are considered as “repeat” customers (they have made more than one transaction), and compare them with the “one-timers”.



The next step was to calculate the number of transactions for each customer and see its distribution. As we can see, it forms a very steep gamma distribution.



One more thing I needed to check, was the lifetime distributions of the customers. For that reason, I calculated the difference between the date of the first transaction, and the date of the last transaction per customer. The resulting distribution is shown below.



At that part, I could move on with the EDA, but I decided to stick to plots that will come at hand when building the model and work as a sanity check during the train and test phase.

2. Model Selection

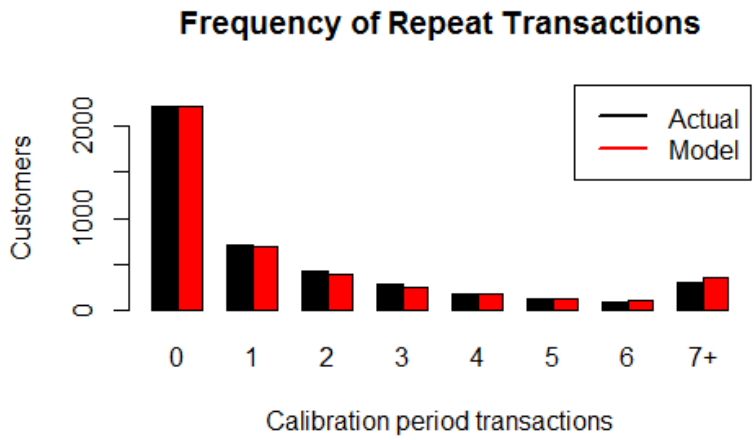
The model I decided to implement is the Pareto/NBD, included in the package “Buy Till You Die” (BTYD). This model is a combination of the Pareto and the negative binomial distribution. The shape of the data needed for the Pareto/NBD model is a 3 column data frame, containing the total transactions of the customer, the date he made his last transaction and the total period that we study the customer, which is the “birth date” (the date of his first transaction) minus the end of the calibration period.

Because the original BTYD package produces some errors when customers made above 100 transactions, I included the script “pnbd.R” which fixes the bugs in the original model.

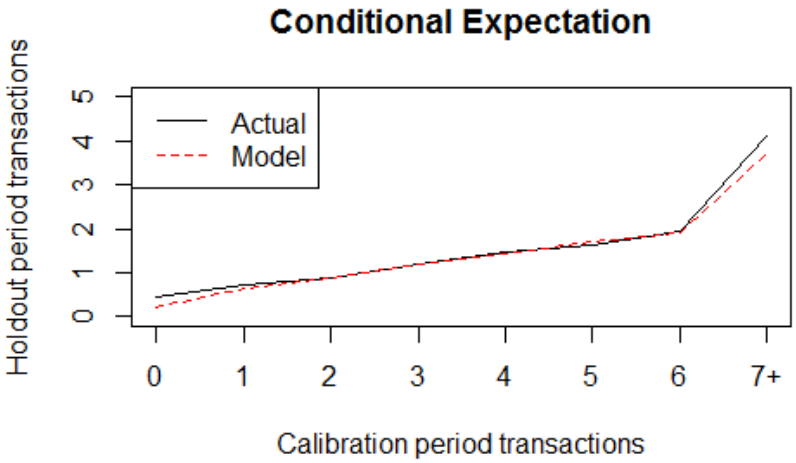
3. Training – Testing – Final Model Process

The language used for the modeling of data is R. The script is called “*Ilias_Katsabalos_Predictions.R*”.

After I transformed the data in the needed shape, I split the set into a 80-20 train and test set, namely in the model as calibration and holdout period. The next step was to estimate the parameters of the model and plot a goodness of fit chart for the calibration period.



After evaluating the fit, I made predictions for the test data and plotted the result. The MSE for the model is **1.290078** compared to a dummy classifier **1.547154**



The final step was to train the model in the whole period and made predictions about the time until the next order. The predictions can be seen in the **results** Data Frame and span across 100 days in the future (max prediction time), which means that if a customer makes a purchase above the 100 limit, the results still display 100. Off course anyone can change this limit according to the needs of the forecasting and general business scenario.

For more information, see the comments on the "*Ilias_Katsabalos_Predictions.R*" file.