**Introduction to Machine Learning – Final Project**

Team 34 – Miki Segall (313562191), Avi Sananes (208057497)

**Executive Summary**

**Data Exploration**

To really understand the data is probably the most important, difficult and exhausting part of a Data Science project. At first, we really wanted to have a better understating and to gain a domain knowledge of the data we are facing, and which features it contains.

In order to do so, we used several data exploration techniques. At first, we wanted to take a glimpse at the train data, so we used the head() function to see how the values in our data look like. At this point we already understood that some serious topics that should be addressed at the pre-processing stage, for example the fact that the duration columns have the word “minutes” in it, and that some of the values are missing.

We noticed that our train data has 10479 rows, which gives us plenty of room for outliers detection and removing, and for separating the dataset to train and validation sets in a later stage. In addition, we saw that we have 22 features (not including the “purchase” column), which we might want to reduce when testing our models.

Afterwards, we used both boxplot and histogram to see the data distribution. We noticed that due to high variation, some of the columns are not really fit the boxplot visualization, but it did help us to realize that some of the columns, like the region column, have obvious outliers. With the histogram visualization, we were able to see that most of the columns are either normally or log-normally distributed. This assumption will serve us for detecting and removing outliers.

Eventually, we used heatmap to plot the feature correlation. We found out that some features, like BoundRates and ExitRates, are strongly correlated, and we might want to consider reducing features with this kind of strong correlation. In addition, we can see that some features are more correlated to “purchase” than other – like PageValues and the “D” column. Once we get to the dimensional reduction part, we will take these conclusions into consideration.

**Pre-processing**

The first part of the pre-processing stage is to normalize the data and modify it in a way that would allow us to use it in machine learning models. At the exploration part of the project, we've reviewed the data manually and looked how the values look like. We understand that we need all the values to be numeric and this required modification.

As the modifications can have several methods of implementation (for example: categorical data can have different encoders), we've created a function that receives a dataframe and which columns should go through which preprocessing. We've determined which processing is required through manually looking at the table and the distinct values of each feature. The preprocessing contains:

* Converting boolean data to 0,1
* Converting month names to numbers
* Extract numbers from string with words = "23.5 minutes"
* Browser data - looking carefully on the values, there is a pattern detected - "<browser\_name>\_<version>" where the version can be in different patterns and styles. We extract the browser name and later encode it, and the browser version we take only the "major" and ignore the "minor" version part.
* Categorical data - encode it with either HotOne or Ordinal encoding. We'll test both of them to check which yields better performance overall, but for the beginning we've set the "HotOne" as default as there is no ordinal connection between the values in the columns so it might come misleading to the model.

**Model implementations and estimations**

**Predictions**

**Summary**

**Appendix – Peer Review**