# Introduction

Finvia is an asset management firm based in Germany. Established in 2019 it is a growing, family-led company with now roughly 100 employees. To best allocate recourses in the marketing and sales department, Finvia needs to predict the number of leads arriving through their website and personal contact via telephone or office visit.

In marketing, a lead is any potential costumer, who comes into a first contact with the company and shows interest (Heinrich, 2020, S. 208). The subsequent steps are acquisition and retention (Kumar & Reinartz, 2018, S. 31). Both steps are recourse intensive. An accurate and secure prediction is therefore very helpful for any asset management firm.

The subsequent paper is structured in five chapters: Firstly, the data will be presented, broadly and with an exploratory data analysis. Afterwards data preparation includes the steps of data cleaning, wrangling, and merging. Then, the modeling will be explained and lastly our results, variable importance, and the performance metrices will be evaluated.

# Data Understanding

The data is stored in four separate tables: *Leads*, *Ads*, *WebsiteTraffic* and *Macro* Data. Every table has a common *Date* column. The beginnings of the tables vary from the 1st of January 2020 (Macro) to the 14th of June 2021 (Ads). All tables last until the 31st of March 2023. The dependent variable *NextDayLeads* is stored in the *Leads* table.

In the table *Ads* numeric values concerning advertising expenditure (*Spend)* and conversion rate (*Impressions* and *Clicks*) are stored. Both variables are split based on platform and funnel. The project is partly based on the assumption that these values have a (lagging) influence on the number of Leads. The analysis will investigate whether spending on advertisement leads to more leads.

The table *WebsiteTraffic* is concerned with the traffic on the company website: [finvia.fo](http://www.finvia.fo). It is measured how many people visit the website and how much time is in total spent on the website. Again, both variables are assumed to have a positive influence on the number of leads by the company. All values until the 15th of August 2021 are equal to zero. Older blog posts on the website suggest a measurement error (<https://www.finvia.fo/wissen/news-feed>).

Macroeconomic data, like the DAX, the electricity and gold price are stored in the table *Macro*. By number of variables this is the biggest table, consisting of 14 variables. It only contains 824 variables in the span of 2958 days. This is because all values are only available for banking days.

## Exploratory Data Analysis

An exploratory data analysis was conducted to better understand the data and its structure. Graphs support this progress.

A provided graph showing *NextDayLeads* over time revealed some extreme outliers, for both the *manual* and the *website* type. Since the clear and timely assignment of leads is difficult (Heinrich, 2020), these outliers suggest measurement errors. Figure 1 shows the weekly sum of *NextDayLeads* after filtering the outliers. This filtering will be described in Chapter 3. Figure 1 clearly suggests that the number of leads is rising over time. This trend shows the increasing importance of this project.

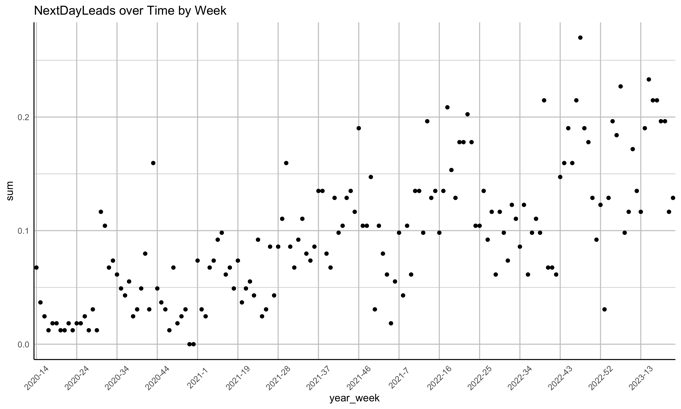


Figure 1: Leads over Time Aggregated by Week, after deleting Outliers.

The analysis of macro-economic data and its relation to leads has found little correlation between the variables. Two variables, the German Electricity price and the US Economy Uncertainty Index stand out. Figure 2 suggests a positive correlation between the German Electricity Price and leads and a negative correlation between the US Economy Uncertainty Index and leads. These correlations are based on the assumptions, that low electricity prices increase fiscal flexibility for consumers and that a low uncertainty gives consumers trust in the financial markets. On the other hand, the colors suggest that these correlations might rather be explained on the time component, which correlates with all three variables (see Figure 1).

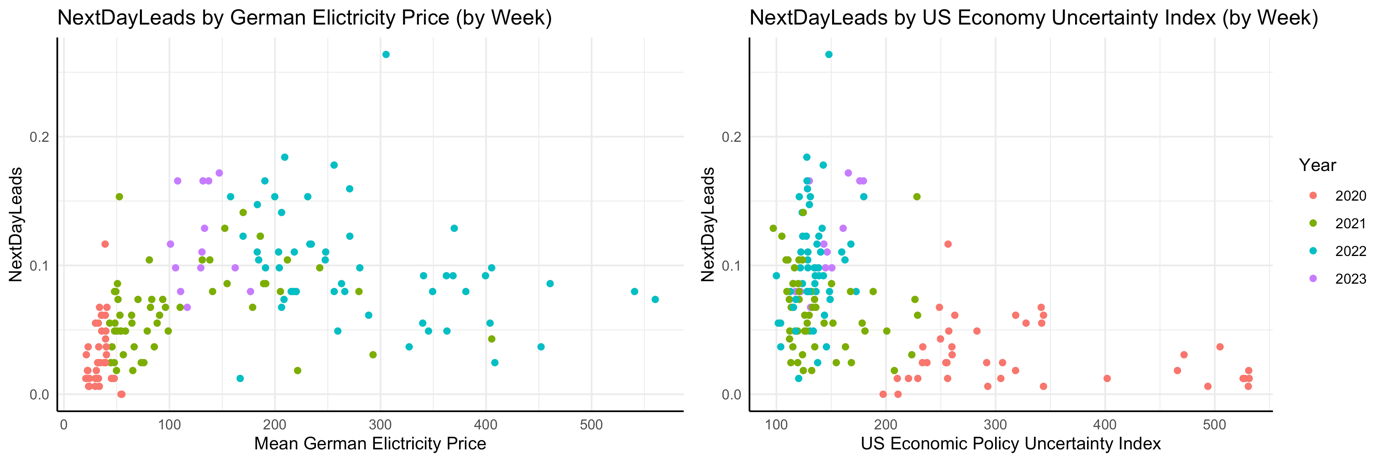


Figure 2: Leads by Week compared to Electricity Price and US Uncertainty Index.

Moreover, relations between the advertisement and traffic data have been examined. While correlations between *Visits* and *TimeSpent* and between *Clicks* and *Impressions* seem to exist, no significant correlation with *NextDayLeads* exists. The highest such correlation is between *Visits* and *NextDayLeads* at around 0.3. Just like the macro-economic data, the code includes graphs to show these values.

# Data Preparation

The goal for data preparation is to have two final data frames, one for manual and one for website leads. These data frames should include only the dependent variable and all the predictors. It should also not include any empty values.

As mentioned in the EDA, the outliers originating from measurement errors are deleted. An outlier is defined as lying four times the standard deviation away from the mean. This combination of mean and standard deviation is a common approach. To take the increasing number of outliers into account, both measures are calculated by year. This way over the whole period 18 observations are filtered out.

All website traffic data prior to the 15th of August 2020 had to be deleted as they display wrong values. Also, for the macro-economic data, the observations from weekends and holidays must be filled out. It was decided to take the values from the previous day, since technically the stock index value, for example, on a Saturday is still the value from Friday.

The columns *Impressions*, *Clicks* and *Spend* from the *Ads* data frame were aggregated by day. All values from all funnels and platforms were summed up to create one single observation per day. Afterwards new features based on the *Ads* and *Traffic* data were created. The *rollapply* function was used to calculate averages over the past days. This function works like a rolling window, adjusting the upper and lower bound for every observation. These mean values were calculated for *Impressions*, *Clicks*, *Spend*, *Visits* and *TimeSpent* and *NextDayLeads*.

Other features created are *month* and *isNextDayWorkDay*. The first stores the month as a factor and functions as a proxy for season. The latter is a binary variable describing whether the next day is a workday or not, meaning weekend or holiday. This variable was created with a specific *holidays* package. The country code for Germany was specified.

To minimize missing data beyond the joining process, an inner join approach was adopted. This strategy entails retaining only those data rows that are present in all data frames. Through this method, the integrity of the dataset is preserved by excluding rows with incomplete information. Keeping all leads observations would have resulted in almost half of observations having incomplete data. This significant proportion harms the modelling process.

Lastly, the data frame was duplicated. In both frames the *Date* column and one *NextDayLeads* was deleted. Both the manual and the website leads now have one data frame. All empty values from the dependent variable were then deleted. They resulted from filtering the outliers and are not useful for prediction. It was important to delete the respective other *NextDayLeads*, as in practice such data is not available and therefore predicting on it, is not feasible. The final data frame has 30 columns and 588 and 583 observations.

# Modeling

# Evaluation