André Bodmer
Krattigstrasse 92
CH-3700 Spiez
+41 77 437 18 55

Data Science Project

Forecasting Call Volume Conceptual Design Report

20th December 2019

ABSTRACT

This report illustrates the conceptual design of the project. The main goal of the project is to forecast the call volume of a contact center (financial industry). These forecasts intend to optimize resource planning; in other words the optimal allocation of staff to the demanded call volume.

First, an investigation of the historical daily time series helps to better understand the structure of the dependent variable (call volume). Second, different features are added to the model to generate the most accurate forecasts. Variables such as weather conditions, central bank press releases, school holidays and public holidays can have an effect on call volumes and will therefore be examined in more detail as explanatory variables.

However, in this conceptual design report no model will be developed to forecast the call volume and no forecasts will be calculated. Rather the required data is first analyzed and discussed with all its facets; automatic extraction, processing, and storage. For this, literature and the contents of the course CAS ADS are consulted and applied. Eventually, the descriptive statistics are calculated and the selected data is analyzed in more detail. Ultimately, this analysis is used to select suitable models and then to calculate forecasts that are as realistic as possible (in Module 3).

TABLE OF CONTENTS

OBJECTIVE

At Valiant's call center personnel planning is currently carried out manually in an Excel spreadsheet. Manual means that the telephone agents are assigned based on the subjective experience of the person responsible for the planning. Regular call patterns or other relevant factors that could have an influence on the call volume are therefore only taken into account "on a gut level" or not at all. Agents are therefore often confronted with unnecessary idle times or an overabundance of telephone calls due to misallocations.

On the other hand, good planning has a direct impact on the customer experience. With enough staff the waiting time for the customer can be reduced[1].
Optimal employee utilization and a positive experience for the customer mean that a correct forecast of the call volume - and based on this an adequate personnel planning - is an indispensable next step towards professionalization of Valiant's call center.

Objectives
This project does not claim to establish a complete personnel planning model. Rather, fundamental considerations are made for a first step, namely the forecast of the call volume. Therefore, the goals are narrowly defined and limited.
This project is to find out

- which features have an influence on the call volume from the telephone channel,
- which modelling technique delivers the best values based on common evaluation metrics,
- to what extent this approach can be implemented in practice and what added value it generates.

The first two objectives can be quantitatively tested using the methods learned in the CAS ADS courses. However, the last goal is then discussed qualitatively.

[1] Call centers are usually measured with a service level of 80/20. This means that at least 80% of all callers must wait less than 20 seconds before being connected to an agent.

Task definition and delimitation

Given the above objective the task in this report is to create a meaningful model for predicting call volume. This is useful in the form of an assessment of the models with regard to the evaluation metrics in the area of supervised learning/regression analysis.

However, this paper does not explicitly claim to develop a complete personnel planning model – this would be the effective demand from the business. Rather, the preliminary work on the implementation of meaningful influencing factors in a forecast model should be prepared. The topics of automated data acquisition at the beginning of the project process and deployment at the end of the project process will be mentioned briefly, but are not the focus of the work.

METHODS

Methodologically, the approach was chosen in close accordance with the CRISP-DM process model. The focus is on the phases “data preparation”, “modeling” and “evaluation”. However, the process steps “business and data understanding” as well as “deployment” are covered in the next chapters in the necessary length.

The raw data set is already balanced and contains only metric data. Thus the necessary descriptive statistics of the individual variables were checked and if necessary transformed (i.e. standardized or normalized). Subsequently, the following models are planned to run.

Models (and other models depending on the content of module 3):

- Classic Methods
    - Linear Regression
    - Linear Ridge Regression
- Machine Learning Methods
    - Decision Tree
    - Support Vector Machine
    - Neural Net
    - Random Forest

R is used for this project. The documents will be commented for metadata. We use mainly functions of the caret-package.

The results obtained are then checked with the learned evaluation metrics and if necessary remodelled, recalculated and re-evaluated until the optimum values are generated.

Evaluation metrics (and other metrics depending on the content of module 3):

- Mean Absolute Percent Error
- (Root) Mean Squared Error
- (Adjusted) R^2
- Mean Absolute Error

Finally, the entire process is carried out using the three common validation techniques in order to ensure that the results are as broadly based as possible and do not lead to incorrect results due to the peculiarities of the respective validation technique.

Validation techniques (depending on the content of module 3):

- Holdout
- Cross
- Bootstrap

Due to the replicability, only the results of the holdout validation are presented in the project report. The results of the cross and bootstrap validation techniques will be given on demand.

With this choice of method, the universe of possibilities is certainly not fully covered, but the most important techniques proposed of Burger (2018), Hastie et al. (2017), James et al. (2017) and Kuhn and Johnson (2013) are applied exemplarily.

DATA

In this chapter all variables are described individually and the corresponding source is denoted. Data automation considerations were clarified with the business in advance. The data comes from various sources: In some cases data can be dragged into the internal and external databases, via web crawlers, or an API. In other cases data comes from automated systems within the bank that are stored in internal databases. In this paper the automated generation of data is no longer discussed. However, it can be said that the possibility of automatic reference was clarified in advance with the business and thus guaranteed. This was an important reason for the selection of the variables listed below, since these variables fulfil this important criterion for automation.

Most data is internal and therefore strictly confidential. For security reasons, internal metric data has been multiplied by a factor only known to the bank.

There are 910 data points available for all the variables shown below. These points were drawn over the period 1 January 2016 to 30 June 2019 (weekends were not taken into account). The descriptive statistics of all variables can be found in Table 1. The density plots can be found in Figure 2 for all variables. In Figure 2 all rows with zero values of the dependent variable (y) have been deleted.

y = Total_Anrufe

The dependent variable (y) to be predicted in this paper is the call volume from the telephone channel. This variable is already automatically stored from the bank's telephone system in the internal databases and can be drawn on a per-minute basis.

Logarithms are not used, since the distribution is already somewhat normally distributed. Rather, the necessary caution is required when interpreting the coefficients.

Source: Valiant

x1 = Pressekonferenzen_EZB

The independent variable (x1) shows the days on which the European Central Bank (ECB) held a press conference [calculated on a daily basis (1= press conference, 0= no press conference)]. The Governing Council takes its monetary policy decision every six weeks, which is explained at the press conference by the President and Vice-President of the ECB immediately after the meeting.

Source: www.ecb.europa.eu

x2 = Lagebeurteilung_SNB

The independent variable (x2) shows the days on which the Swiss National Bank (SNB) published a situation assessment [calculated on a daily basis (1= situation assessment, 0= no situation assessment)]. In the months of March, June, September and December, the SNB conducts an in-depth monetary policy assessment, which results in a monetary policy decision and the publication of a medium-term, conditional inflation forecast. The National Bank justifies its decisions in a press release and in the quarterly report on monetary policy, which is published in the Quarterly Bulletin.

Source: www.snb.ch

x3 = Schulferien_Aarau

The independent variable (x3) represents the holidays of the canton of Aarau (daily basis: 1= holiday day, 0= no holiday day). This data comes from the database of the Swiss Conference of Cantonal Ministers of Education (EDK), which can be accessed via public holiday calendars.

Valiant has a particularly strong presence in the regions of Bern, Lucerne and Aargau, which is why the school holidays of these cantons (cities) are used.

Source: www.edk.ch

x4 = Schulferien_Bern

The independent variable (x4) represents the holidays of the Canton of Berne (daily basis: 1=holiday day, 0=no holiday day). This data comes from the database of the Swiss Conference of Cantonal Ministers of Education (EDK), which can be accessed via public holiday calendars.

Valiant has a particularly strong presence in the regions of Bern, Lucerne and Aargau, which is why the school holidays of these cantons (cities) are used.

Source: www.edk.ch

x5 = Schulferien_Luzern

The independent variable (x5) represents the holidays of the canton of Lucerne (daily basis: 1= holiday day, 0= no holiday day). This data comes from the database of the Swiss Conference of Cantonal Ministers of Education (EDK), which can be accessed via public holiday calendars.

Valiant has a particularly strong presence in the regions of Bern, Lucerne and Aargau, which is why the school holidays of these cantons (cities) are used.

Source: www.edk.ch

x6 = Nationale_Feiertage

The independent variable (x6) shows the days on which a national holiday took place [on a daily basis (1= national holiday, 0= no national holiday)]. This variable is taken from the internal Facility Management Report.

Source: Valiant

x7 = Imagekampagne

The independent variable (x7) shows the days on which Valiant launched a national image campaign [on a daily basis (1= National image campaign launched, 0= No national image campaign launched)]. This variable is taken from the Bank's internal marketing reporting, which is maintained manually but can be automatically extracted from the database.

Source: Valiant

x8 = Medienmitteilung

The independent variable (x8) shows the days on which Valiant published a media release [on a daily basis (1= National media release published, 0= No national media release published)]. Those press releases were selected which have a holistic character for the bank. Regional news were omitted due to local influence. Instead, quarterly, half-yearly and annual results as well as strategy announcements and information for the Annual General Meeting were marked as relevant media releases. This variable originates from the Bank's internal communication reporting, which is maintained manually but can be extracted automatically from the database.

Source: www.valiant.ch/news

x9 = SMI

The independent variable (x9) shows the daily closing value of the SMI Index. The variable comes from the Bank's internal databases (Product Management and Operations). Logarithms are not used because the distribution is already reasonably normal. Rather, the necessary caution is required when interpreting the coefficients.

Source: Valiant / Bloomberg

x10 = SPI

The independent variable (x10) shows the daily closing value of the SPI Index. The variable comes from the bank's internal databases (Product Management and Operations). Logarithms are not used because the distribution is already reasonably normal. Rather, the necessary caution is required when interpreting the coefficients.

Source: Valiant / Bloomberg

x11 = Kurs_Valiant_Aktie

The independent variable (x11) shows the daily closing value of the Valiant share. The variable is taken from the Bank's internal databases (Product Management and Operations). Logarithms are not used because the distribution is already reasonably normal. Rather, the necessary caution is required when interpreting the coefficients.

Source: Valiant / Bloomberg

x12 = Celsius

The independent variable (x12) shows the air temperature in degrees Celsius (daily average). Valiant has a particularly strong presence in the regions of Aargau, Bern and Lucerne, which is why the weather data from these cantons (cities) are used. Therefore, only the data from the weather stations in these regions were selected and the arithmetic average drawn. The variable originates from the archive data of MeteoSwiss' ground monitoring networks, which thanks to IDAWEB provide interactive data. The data archive contains meteorological measurement and observation data of Switzerland from the beginning of the measurement up to the current day before. The data are supplied as text files (ASCII) and zipped. The formats CSV (semicolon separated) and Bulletin (fixed column width) are available.
Source: www.idaweb.ch

x13 = mm

The independent variable (x13) shows the precipitation in mm (daily sum 24h). Valiant is particularly well represented in the regions of Aargau, Bern and Lucerne, which is why the weather data of these cantons (cities) are used. Therefore, only the data from the weather stations in these regions were selected and the arithmetic average drawn. The variable originates from the archive data of MeteoSwiss' ground monitoring networks, which thanks to IDAWEB provide interactive data. The data archive contains meteorological measurement and observation data of Switzerland from the beginning of the measurement up to the current day before. The data are supplied as text files (ASCII) and zipped. The formats CSV (semicolon separated) and Bulletin (fixed column width) are available.
Source: www.idaweb.ch

x14 = h

The independent variable (x14) shows the sunshine duration in h (daily total). Valiant is particularly well represented in the regions of Aargau, Bern and Lucerne, which is why the weather data of these cantons (cities) are used. Therefore, only the weather station data from these regions were selected and the arithmetic average was drawn. The variable originates from the archive data of MeteoSwiss' ground monitoring networks, which thanks to IDAWEB provide interactive data. The data archive contains meteorological measurement and observation data of Switzerland from the beginning of the measurement up to the current day before. The data are supplied as text files (ASCII) and zipped. The formats CSV (semicolon separated) and Bulletin (fixed column width) are available.
Source: www.idaweb.ch

Table 1 shows the values of common descriptive statistics and other relevant information on all variables. A histogram, the density plot, and the box plot for each variable will be provided on demand.

Table 1: Descriptive Statistics

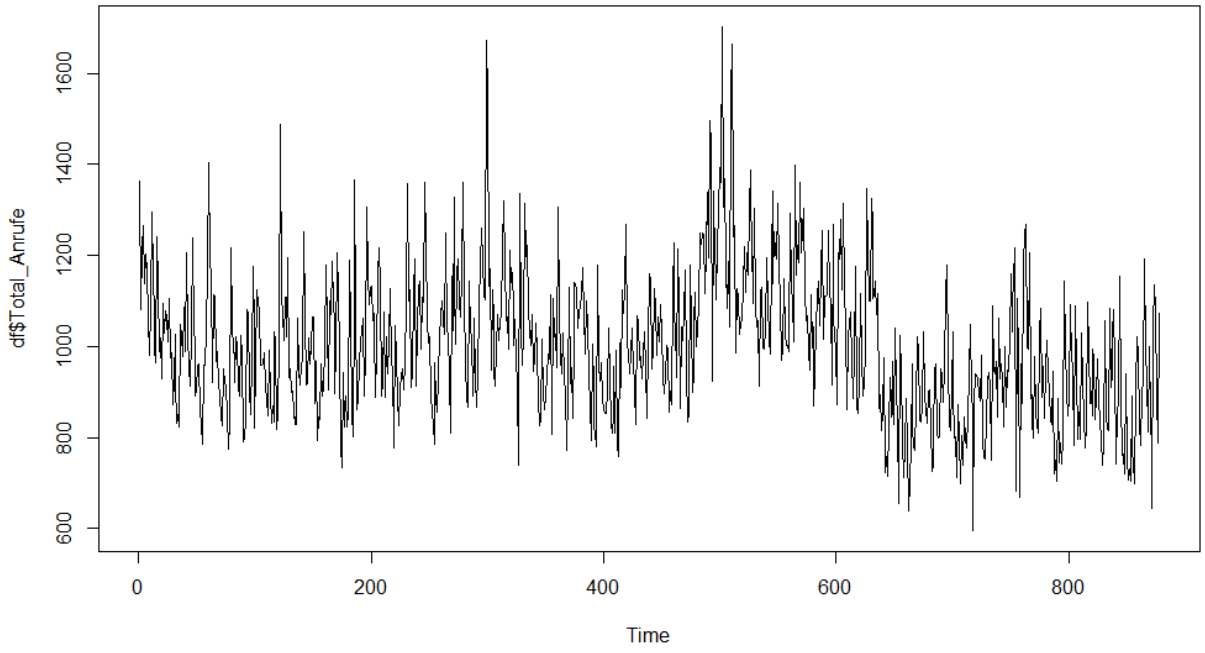
Variable	N	Art	Min.	1st Q.	2nd Q.	3rd Q.	Max.	Mean	SD	Skew.	Kurt.
Total_Anrufe	*910*	*Metric*	*0*	*876*	*981*	*1094*	1703	*965*	*240*	*-1.938*	*10.046*
Pressekonferenzen_EZB	910	Metric Binary	0	0	0	0	1	0.038	0.173	5.434	30.532
Lagebeur-teilung_SNB	910	Metric Binary	0	0	0	0	1	0.015	0.123	7.875	63.016
Schulferien_Aarau	910	Metric Binary	0	0	0	1	1	0.254	0.435	1.131	2.280
Schulferien_Bern	910	Metric Binary	0	0	0	1	1	0.263	0.440	1.079	2.164
Schulferien_Luzern	910	Metric Binary	0	0	0	1	1	0.285	0.451	0.955	1.911
Nationale_Feiertage	910	Metric Binary	0	0	0	0	1	0.032	0.176	5.330	29.412
Image-kampagne	910	Metric Binary	0	0	0	0	1	0.192	0.394	1.561	3.438
Medien-mitteilung	910	Metric Binary	0	0	0	0	1	0.037	0.190	4.879	24.804
SMI	910	Metric	7497	8234	8826	9087	9979	8729	547	-0.097	2.183
SPI	910	Metric	7783	8943	10232	10638	12054	9939	978	-0.189	2.045
Kurs_Valiant_Aktie	910	Metric	88	103	107	112	120	107	6.725	-0.572	2.863
Celsius	910	Metric	-9.70	3.50	8.90	15.60	26.70	9.46	7.618	0.054	2.116
mm	910	Metric	0	9.82	22.45	38.17	94.20	26.58	20.717	0.825	3.096
h	910	Metric	0	1.20	4.50	8.80	14.50	5.16	4.273	0.440	1.926

Source: Own presentation

The analysis of descriptive statistics shows that all variables are already metric. Many of them are binary, i.e. they have either the value 0 or 1. Due to the properties of the individual variables no transformations in the form of standardization or normalization are performed for the time being. The data set can be reused accordingly.

Figure 1 shows the structure of the dependent variable. All those points which contain a zero value (public holidays) are deleted.

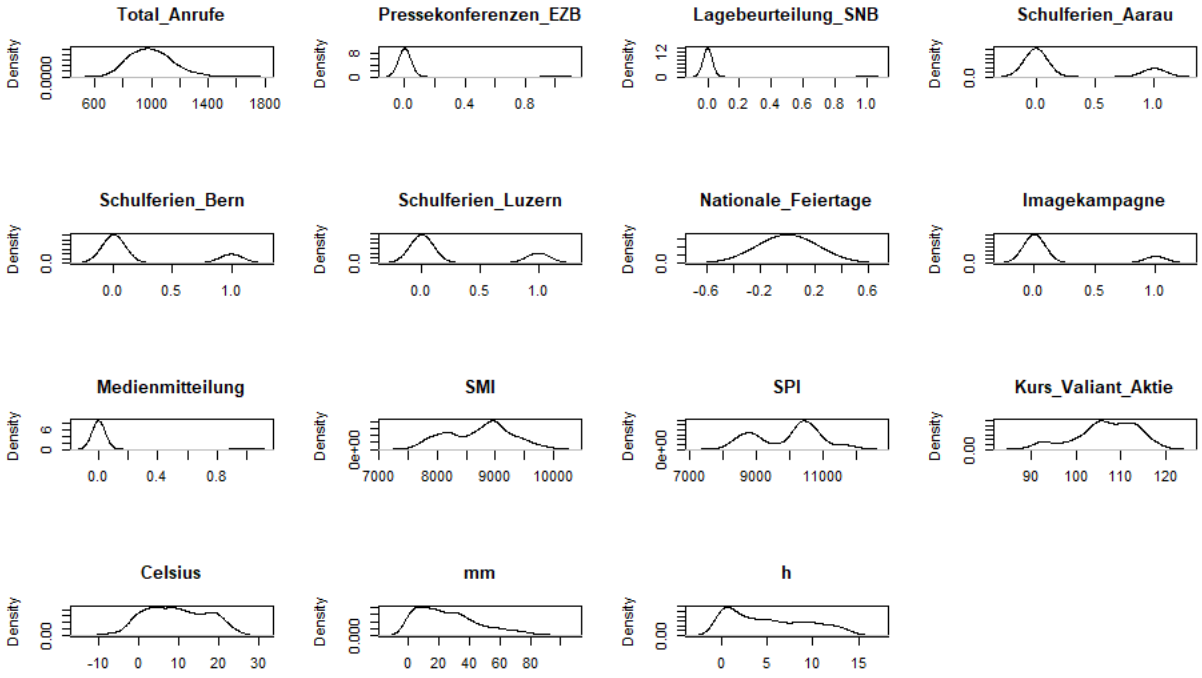
Figure 1: Call volume by day



Source: Own presentation

Figure 2 shows the density plots for all variables.

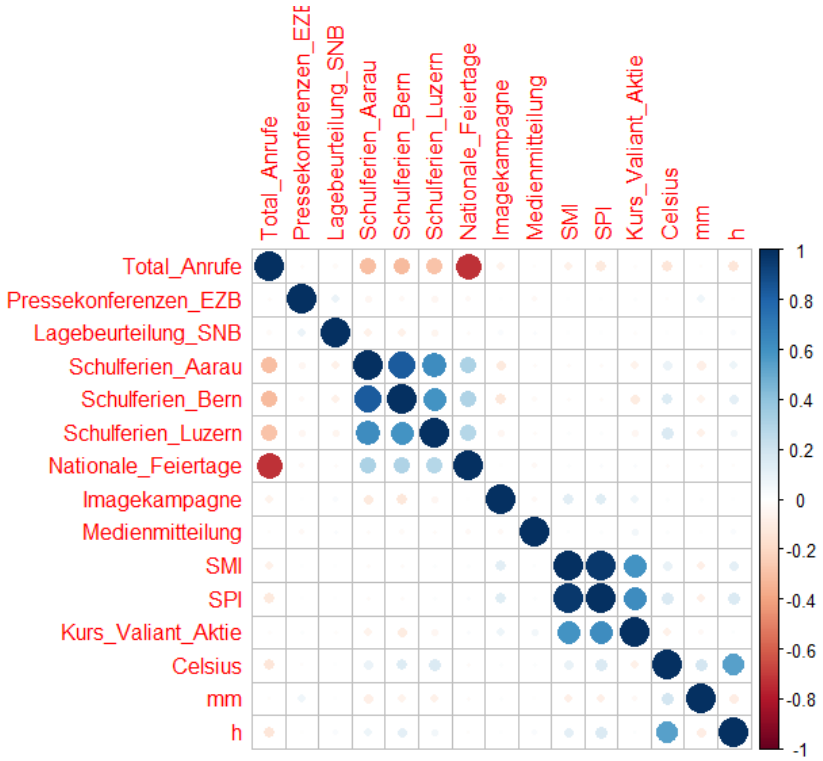
Figure 2: Densitiy plots for all variables



Source: Own presentation

As a further step, the correlations of the individual variables are considered in Figure 3. This is done because an initial overview of a possible dependency (not causality) between the x variables and y can be detected. In addition, many regression methods only work under the assumption that the x-variables are independent of each other (correspondingly also the naming of the variables).

Figure 3: Correlogram



Source: Own presentation

METADATA

There is a lot of additional information of the data in this project that can be considered as metadata. Some additional information has already been given in the previous chapter regarding the description of the data. Thus the sources of the individual variables were indicated or measures of the variables were formulated.
For example, the daily closing value was used for the SMI, SPI and the Valiant share price. This additional information must be added as metadata. Furthermore, the variable media release only takes into account those that have a national character. The weather variables were calculated as averages over several cantonal measuring weather stations.
Finally, for the dependent variable y, all weekends and holidays were excluded. Otherwise we would have zero values in the data set.
If this information is not described in the notebook as markdown text, one could directly append the meta information to the variables in the code with ".myinstrument_name" in case we work with Python.
But for these first analyses I added the metadata as comment in the R script.
As I use R Studio metadata was written as #-comments. Hence, all people who have access to these files can reproduce every single step.
Accordingly, two R scripts were also created: the first script is responsible for preprocessing. There are the derivation shown why single variables were used or rejected resp. standardized. In the second script, the modified data set was only used for modelling. In this script only metadata concerning the parameterization is added.

DATA QUALITY

In this project methods from the field of supervised learning are applied primarily. Thus, mainly regression methods are used for forecasting, the corresponding evaluation metrics from this area are also used. Burger (2018), Hastie et al. (2017), James et al. (2017) and Kuhn and Johnson (2013) recommend Mean Squared Error, Mean Absolute Percent Error, Mean Absolute Error and (Adjusted) R^2 for this type of modelling. The first three values should be as small as possible. This means that the errors of the model must be minimized. The last value should be as large as possible, whereby it could be "artificially" adjusted and the problem of overfitting can occur. As benchmark we want to reach a Mean Absolute Percent Error below 0.1 and a R^2 above 80%.

In R exist the step()-function which automatically removes variable with no impact on the depended variable. The caret-package also has functions that remove variables that are highly correlated.

With these methods, it is tried to eliminate qualitatively bad data. Nevertheless, the quality of data can be problematic. Especially the weather data are critical: For the time being it is an average of regional measuring stations. For example, if you look at the weather in the city of Bern, the city of Biel, and the city of Thun – all three cities are located in the canton of Bern – the weather can be completely different due to regional conditions. Therefore, these variables are considered with particular skepticism.

DATA FLOW

How we download and store the data was mentioned in the chapter "data". At Valiant, there is a central data department that stores all data on dedicated SQL databases. There is an integration layer, which standardizes the data obtained via various interfaces. For machine learning projects, Dataiku is used in the first place. This program is used for preprocessing, modelling and storing the outputs.
Since R and Python are closer to the author, R is used for the time being in this project. In module 3 of the CAS ADS the author wants to work with Python. In module 5, I also complete the Dataiku course in order to respond to the bank's wishes. To visualize the results and the data Valiant uses Power BI of Microsoft.

Eventually, it was agreed with the business that the final model would be implemented in our data warehouse and applied to our SQL server. Since this project was carried out with R, the corresponding operators are also used to read and write PMML models. The models are then saved as XML files and can be reused. Valiant works with Dataiku for machine learning projects. Accordingly, this transition from R to Dataiku using PMML models is necessary.
The automation of data selection is then institutionalized and the model is enriched with new data every day. If necessary, the productive model has to be adjusted. This final model for the call volume must then be extended with another model, namely a translation table to derive the personnel requirements based on the call volume. However, this model is not part of this project.

DATA MODELS

Logical

All data is downloaded on infrastructure of the bank. These infrastructure is in the bank environment and on premise. Moreover, R is used for this conceptual work. The package caret helps to do the work in this project with less code. Nonetheless, the conventional way with non-performant code typing was chosen as well. The reason for this procedure is to better understand each individual step given the high-level language of caret.
There are two R scripts. The first is to preprocess data and the second to run the models. The versions are labeled with two numbers: X.Y. X is the version and Y is the subversion. For every major release the version changes, for every minor release the subversion changes.

Physical

This project I have done on my personal Laptop (HP, intel core i7 vPro, 8 MB RAM) and can be done with every consumer laptop as well.

Conceptual

The following models are now carried out using the holdout validation as an example:

- Classic methods
    - Linear Regression (lm)
    - Linear Ridge Regression (lmridge)
- Machine Learning Methods
    - Decision Tree (rpart)
    - Support Vector Machine (svm)
    - Neural Net (nnet)
    - Random Forest (randomForest)

RISKS

There are several risks in this project. For example, the telephone software was converted in June 2018; a new conversion will follow in 2020. As a result of these conversions, it is not always possible to measure the same amount. For this reason, however, an attempt is made to bring the measurements up to the same level with a translation table after each changeover so that comparability can be guaranteed.
There are also many projects in progress at the bank that have a direct influence on the call volume, which is not represented in any of the variables mentioned. However, this absolute isolation from the influence of individual variables on the call volume will always be an issue. Accordingly, it must be tried to make the error as small as possible with the explanatory variables and to predict the explained variable as well as possible.
Problems with the versioning or the storage space on the servers of Valiant don't seem to be a big risk at the moment, because we can't talk about a big data project in this project yet.

However, if the data quality of the selected data turns out to be poor, then the project must work with alternative variables or use the same variables from alternative providers (e.g. weather data).

In these fast moving times, new packages and versions of the software R and Python come onto the market. The monitoring of new versions is guaranteed by the bank. However, code often has to be rewritten so that the scripts run error-free again. When calculating the project resources, a certain number of man-days are certainly included for the running operation.

PRELIMINARY STUDIES

First, the presented results were run with the data set without public holidays. With the step()-function, this model has filtered out the image campaign, the media releases and the variables celsius and mm. Depending on the generation of the data set (train and test via holdout technique), the deleted variables may differ.

Figure 4: Output linear regression

Call:
lm(formula = Total_Anrufe ~ Pressekonferenzen_EZB + Lagebeurteilung_SNB +
Schulferien_Bern + Schulferien_Luzern + SMI + SPI + h, data = train)

Residuals:

Min	1Q	Median	3Q	Max
-373.9	-104.2	-17.7	92.5	638.7

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	596.3252	166.6878	3.58	0.00038	***
Pressekonferenzen_EZB	-70.7672	35.1283	-2.01	0.04447	*
Lagebeurteilung_SNB	-101.2364	61.9810	-1.63	0.10300	
Schulferien_Bern	-27.9725	18.5453	-1.51	0.13208	
Schulferien_Luzern	-30.5778	17.4286	-1.75	0.07994	.
SMI	0.1813	0.0458	3.95	8.7e-05	***
SPI	-0.1130	0.0259	-4.36	1.6e-05	***
h	-5.9460	1.5633	-3.80	0.00016	***

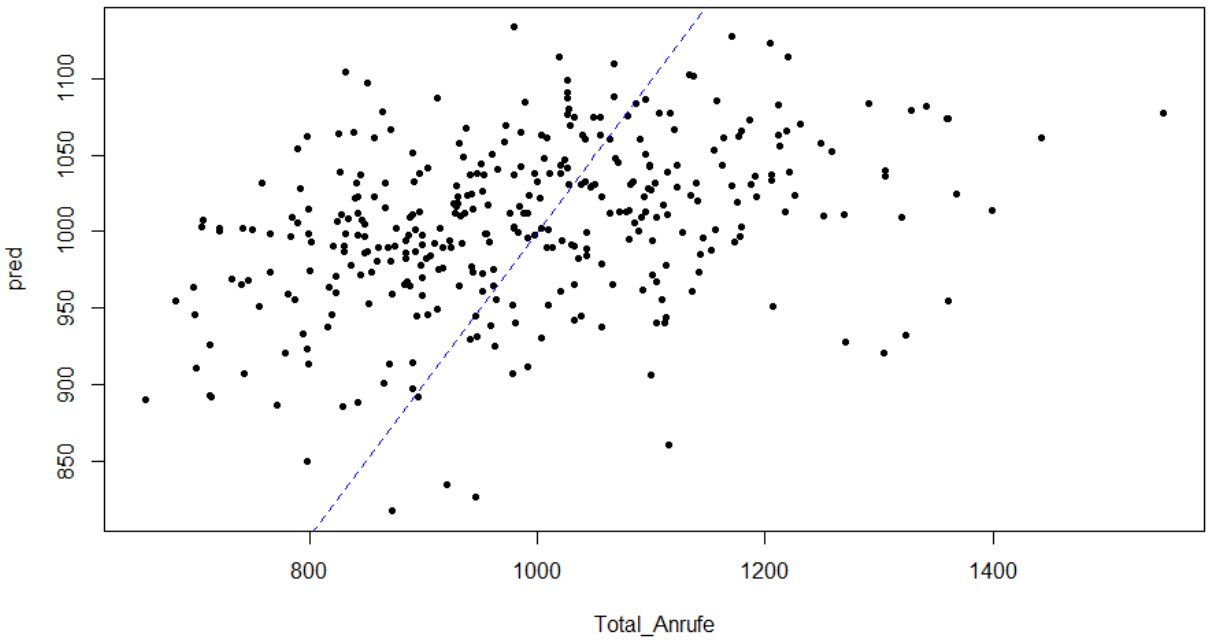
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 150 on 522 degrees of freedom
Multiple R-squared: 0.107, Adjusted R-squared: 0.0955
F-statistic: 8.98 on 7 and 522 DF, p-value: 1.8e-10

Source: Own presentation

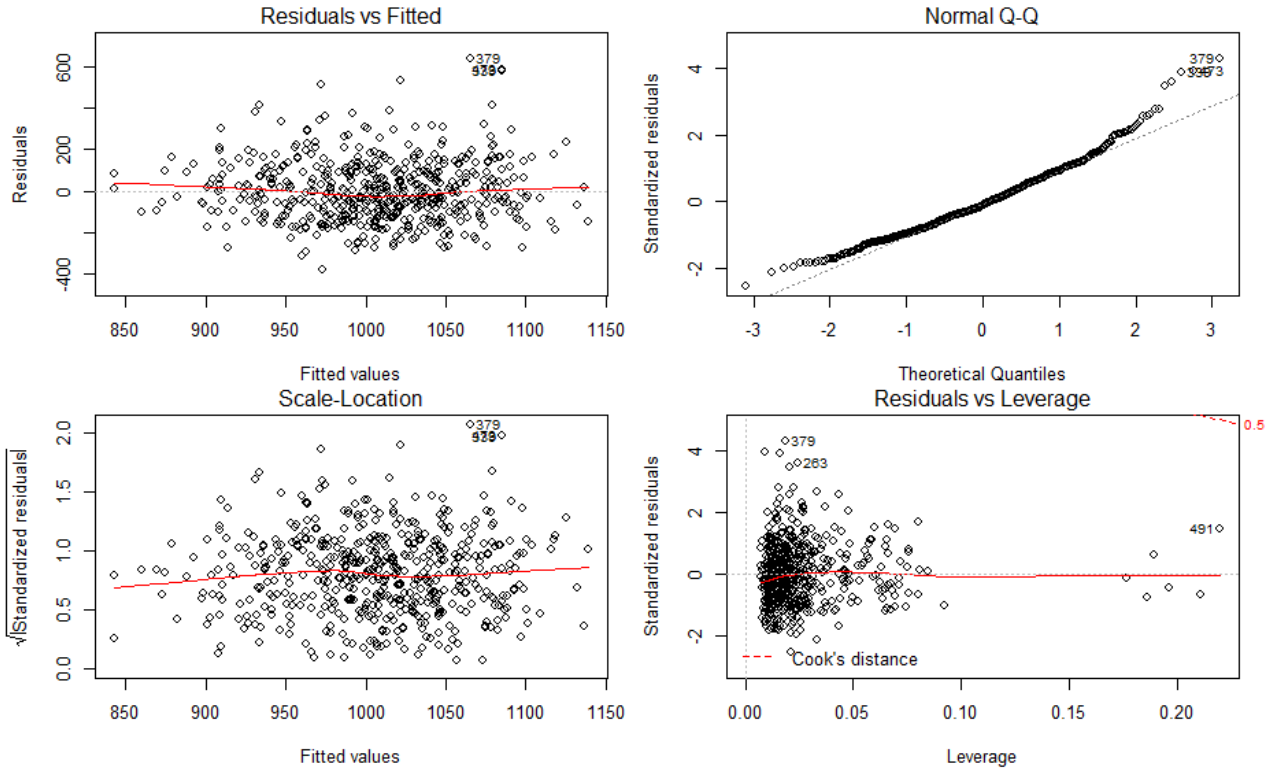
Obviously, the variable h and the indices seem to have a significant influence on call volume.

Figure 5: Linear regression, test vs. prediction



Source: Own presentation

Figure 6: Linear Regression, errors



Source: Own presentation

If we perform the ridge regression and increase lambda sequentially with the values 0 to 1 in 0.1 steps, we get the following result. The variables that have an influence on y are identical to the standard regression model.

Figure 7: Output linear ridge regression

Call:

lmridge.default(formula = form, data = train, lambda = seq(0, 1, 0.1), contrasts = NULL)

Coefficients: for Ridge parameter K= 0

	Estimate	Estimate (Sc)	StdErr (Sc)	t-value (Sc)	Pr(>|t|)	
Intercept	6.084e+02	5.237e+06	8.002e+06	0.654	0.513	
Pressekonferenzen_EZB	-6.878e+01	-2.944e+02	1.515e+02	-1.943	0.053	.
Lagebeurteilung_SNB	-9.816e+01	-2.391e+02	1.526e+02	-1.567	0.118	
Schulferien_Aarau	2.064e+01	1.908e+02	2.713e+02	0.703	0.482	
Schulferien_Bern	-4.159e+01	-3.971e+02	2.607e+02	-1.523	0.128	
Schulferien_Luzern	-3.249e+01	-3.259e+02	1.885e+02	-1.729	0.084	.
Imagekampagne	-1.595e+01	-1.464e+02	1.536e+02	-0.953	0.341	
Medienmitteilung	-1.824e+01	-8.003e+01	1.521e+02	-0.526	0.599	
SMI	1.700e-01	2.134e+03	5.899e+02	3.618	<2e-16	***
SPI	-1.080e-01	-2.406e+03	6.195e+02	-3.884	<2e-16	***
Kurs_Valiant_Aktie	4.390e-01	7.053e+01	2.090e+02	0.338	0.736	
Celsius	-8.430e-01	-1.498e+02	2.083e+02	-0.719	0.472	
mm	-1.870e-01	-8.998e+01	1.635e+02	-0.550	0.582	
h	-5.240e+00	-5.204e+02	1.878e+02	-2.771	0.006	**

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Ridge Summary

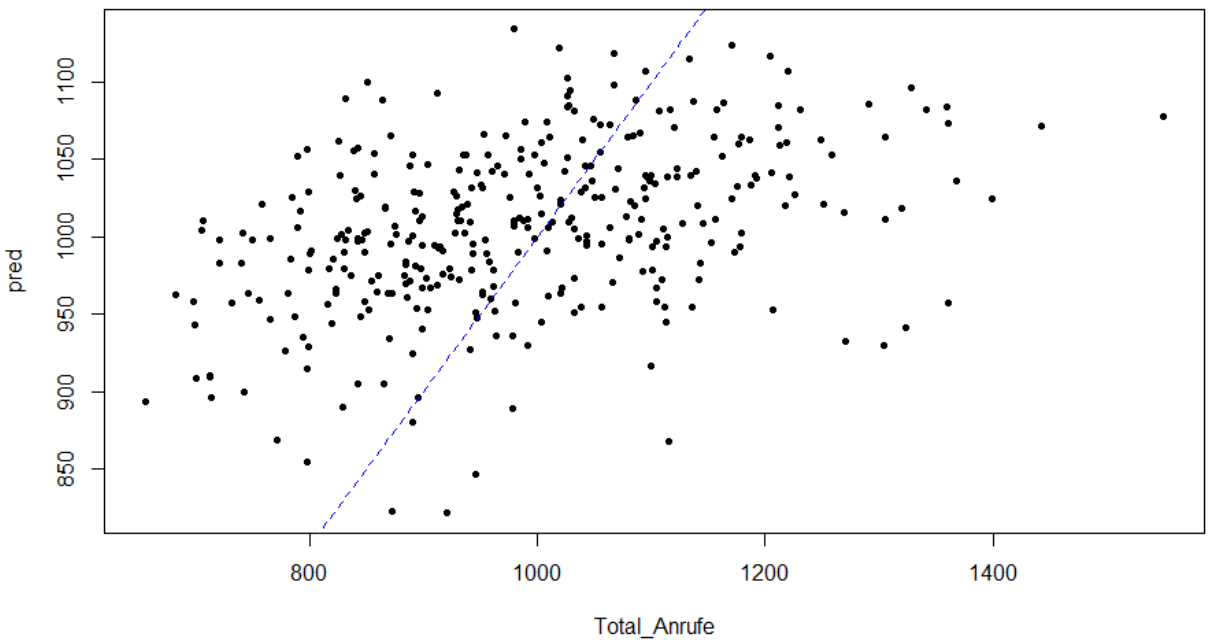
R2	adj-R2	DF ridge	F	AIC	BIC
0.11340	0.09280	12.99996	5.08553	5323.90455	8704.07660

Ridge minimum MSE= 1150810 at K= 0

P-value for F-test (13 , 517) = 1.60425e-08

Source: Own presentation

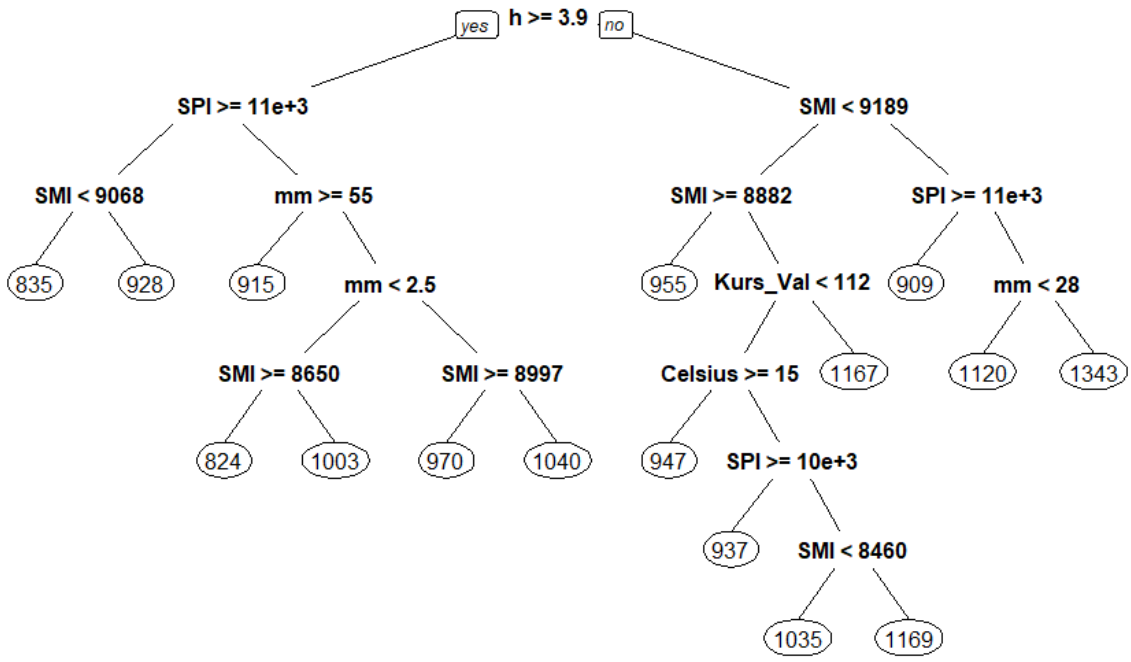
Figure 8: Linear ridge regression, test vs. prediction



Source: Own presentation

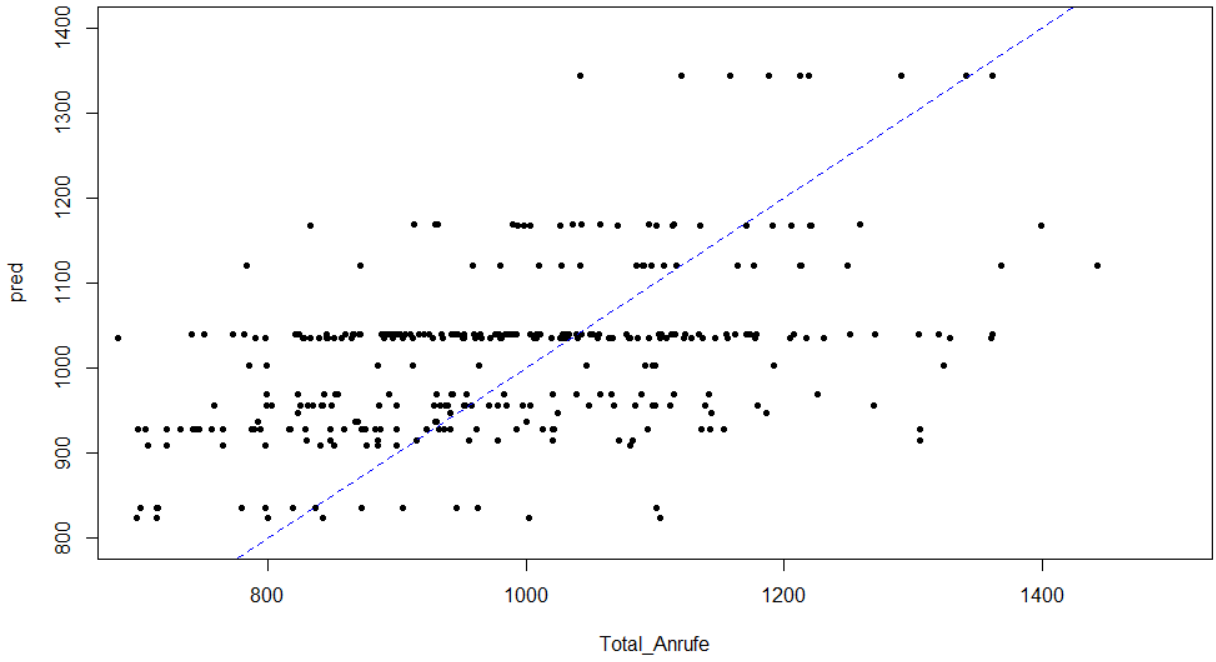
If now more modern methods are applied with default parameters, we get the following results. First the results of the decision tree model.

Figure 9: Output decision tree



Source: Own presentation

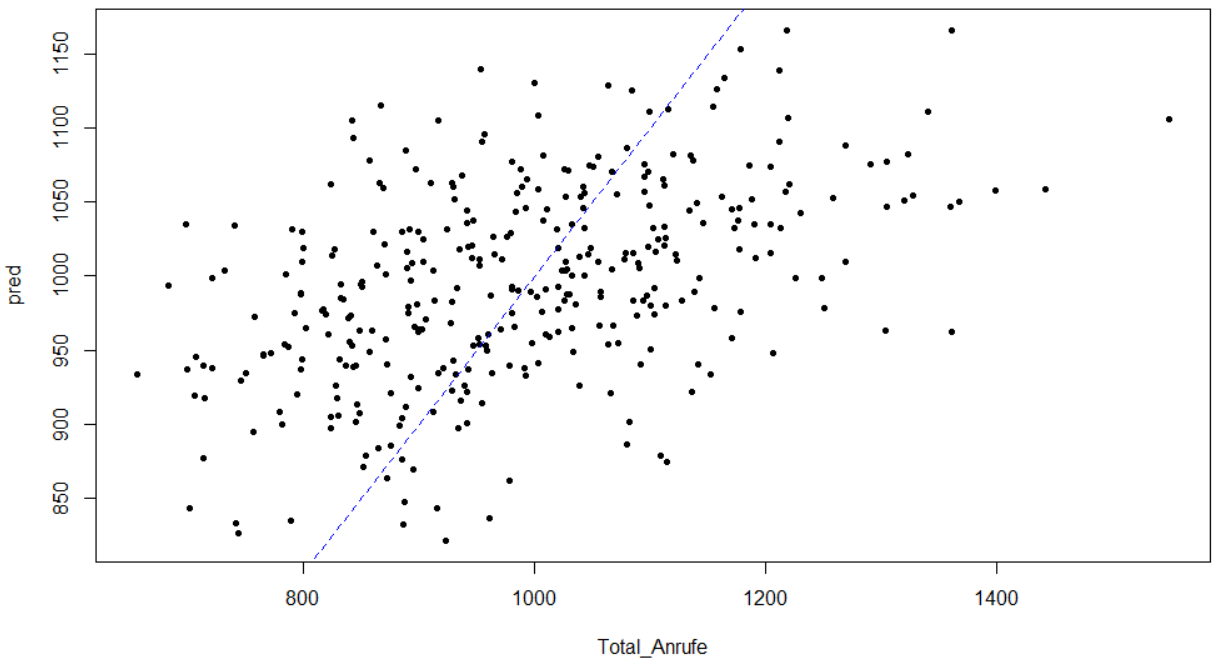
Figure 10: Decision tree, test vs. prediction



Source: Own presentation

Here are the results with the support vector machine model.

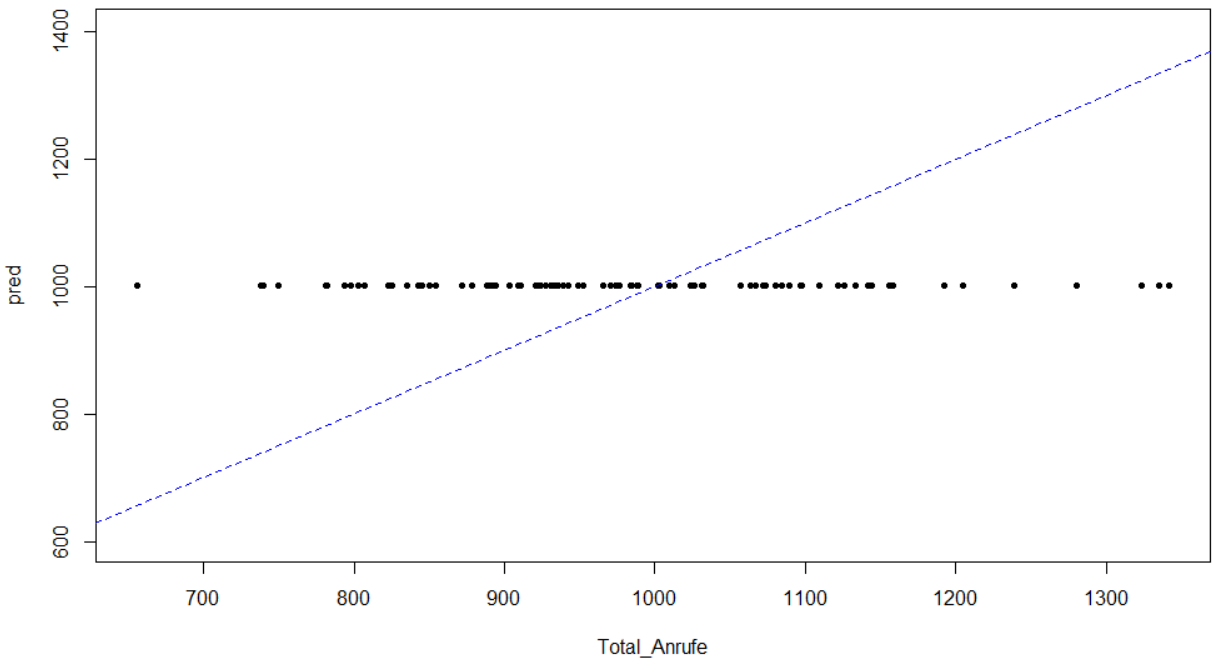
Figure 11: Support vector machines, test vs. prediction



Source: Own presentation

Here are the results with the neural net with one hidden layer.

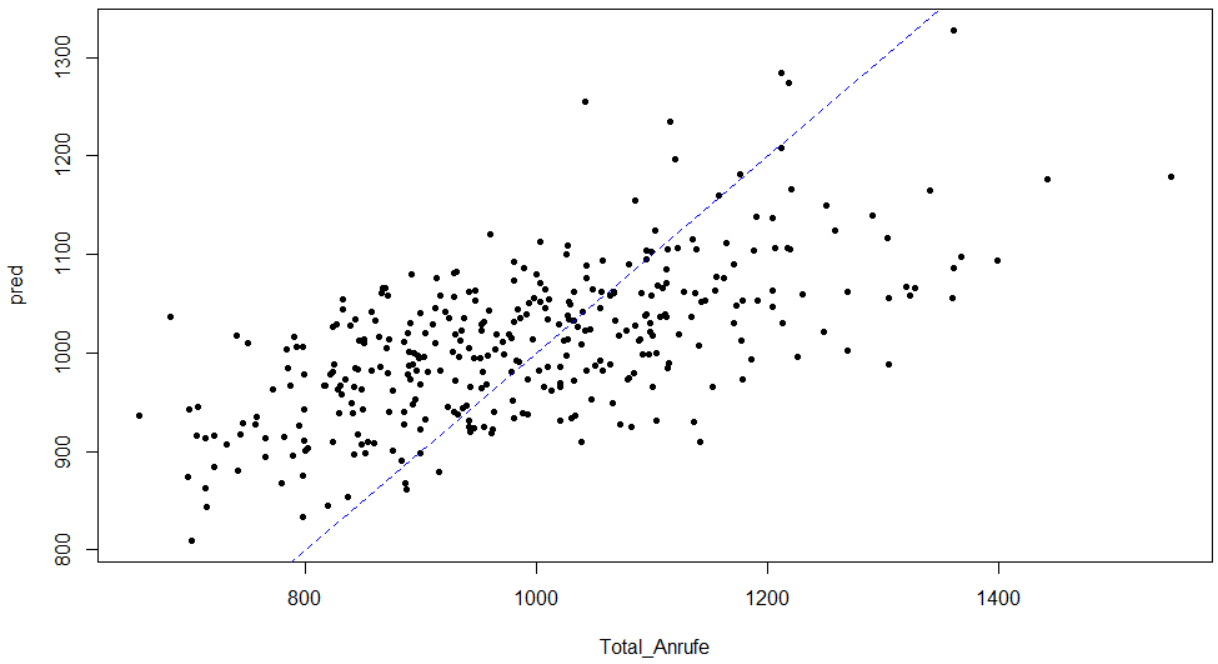
Figure 12: Neuronal network, test vs. prediction



Source: Own presentation

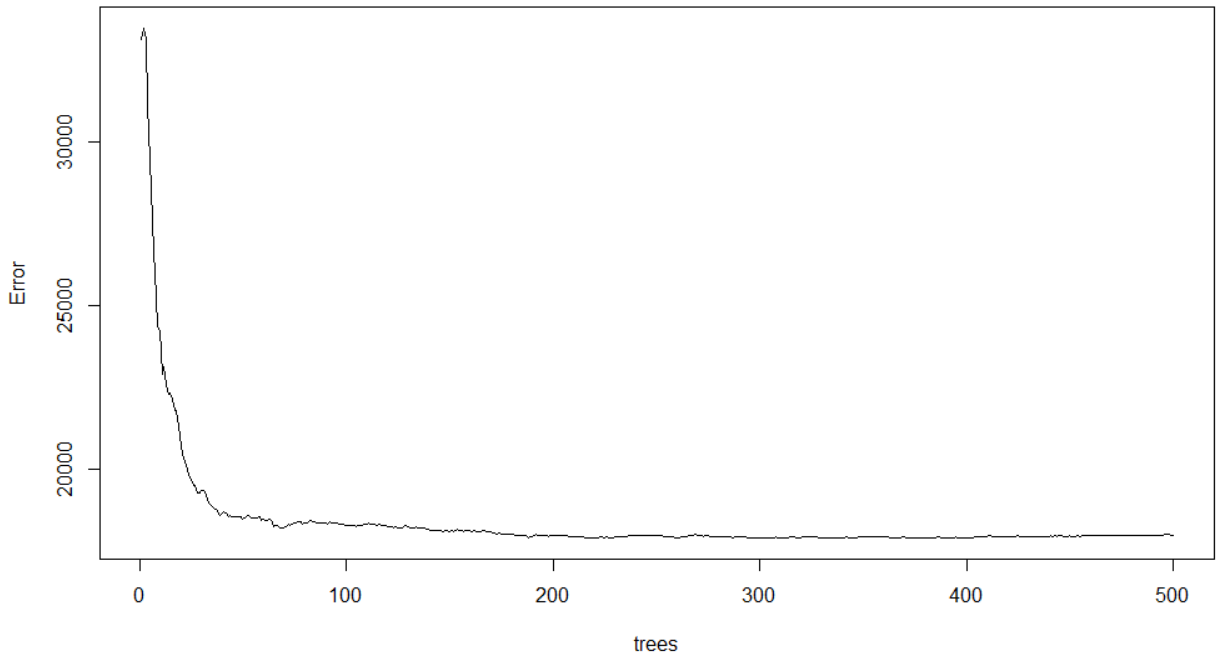
And eventually, the random forest method:

Figure 13: Random forest, test vs. prediction



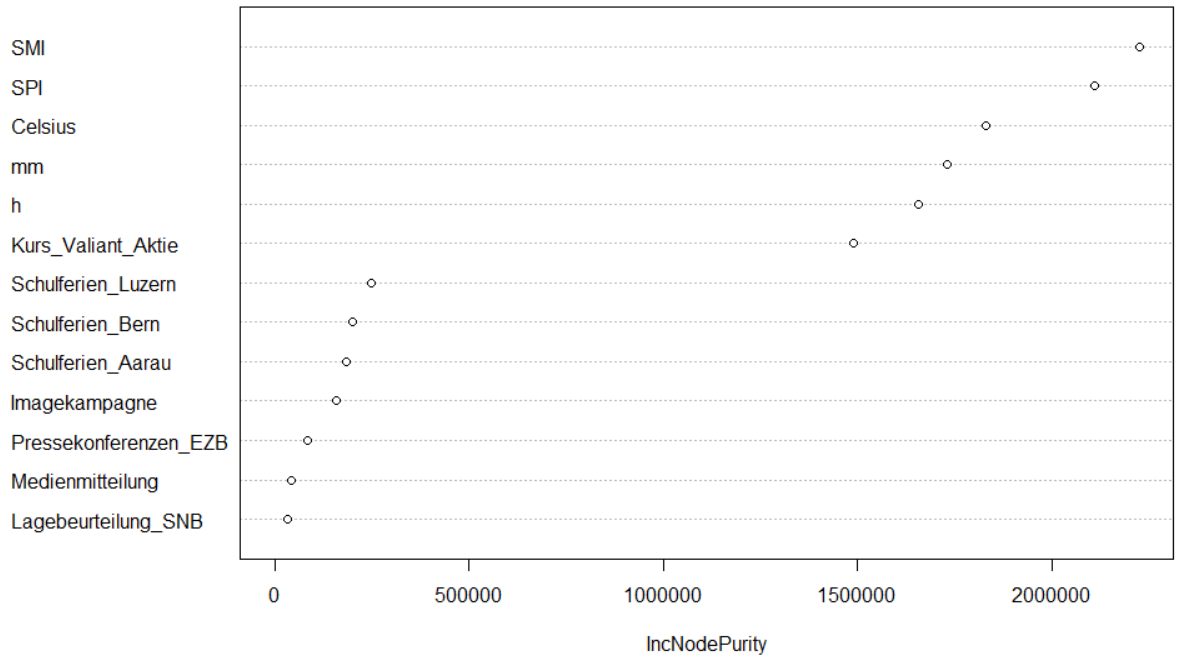
Source: Own presentation

Figure 14: Random forest, errors vs. trees



Source: Own presentation

Figure 15: Random forest, variable-importance-plot



Source: Own presentation

Table 2 shows the evaluation metrics of the models.

Table 2: Evaluation metrics

Categories	MAPE	RMSE	R^2	MAE
Linear Regression (lm)	0.116	144	0.140	116
Linear Ridge Regression (lmridge)	0.115	142	0.169	115
Decision Tree (rpart)	0.112	141	0.222	113
Support Vector Machine (svm)	0.109	137	0.222	109
Neural Net (nnet)	0.126	155	NA	127
Random Forest (randomForest)	0.098	123	0.407	99

Source: Own presentation

In this variant, the model with the random forest method shows the best results. The results for the cross- and bootstrap-validation techniques will be provided on demand. With these techniques the random forest models also provide the best results.

The random forest models provide the smallest error values or the largest R^2 for all validation techniques. This result is accepted for this project. But our goal was a MAPE value under 0.1 (fulfilled) and a R^2 bigger than 0.8 (not fulfilled). For further practice, different hyperparameters are experimented with, as some model specifications would allow this. To depict all the variations in this project report would go beyond the scope. But in practice this is a necessary step. Accordingly, this temporary result must be treated with caution. The regression with higher polynomial equations was also not performed as an extension. Models which take into account

the frequency structures of the call volume would ideally have to be applied. Time series models, such as ARMA or RNN models, would probably have achieved better values. However, these time series models will be applied for the finalization of the model.
Other features can also be included in addition to this existing set: For example, the varying call duration could be important for personnel planning. The analysis of the call groups could also be important, because, for example, after an interest rate adjustment by the SNB, not the same people do call as in a national image campaign by Valiant.

Nevertheless, this work was able to provide an answer to the question as to which features assumed by the business to be significantly explanatory for the call volume are. These are school holidays, national holidays and sunshine hours. Further features can be integrated optionally.
It was also possible to achieve the goal of which model provides the best values expressed in evaluation metrics – although these results are regarded as an interim result.
And finally, we come to our final goal: To what extent can the preliminary work created in this project be integrated into a productive Valiant framework? This goal was discussed with the business and judged to have been achieved. The only disadvantage is that we initially worked with R and not directly with Dataiku. However, due to the PMML models implemented at Valiant – or the same standards – this transfer can still be carried out.
The division of the R scripts into two had proven to be successful. In other words, a script was created which prepares the data and transforms any variables. The other script calculates and evaluates the models. A third script to prepare the final model for deployment has not yet been created.
Finally, it can be said that whenever possible, the frequency should be increased to intraday. On the one hand, more data points are available and personnel planning can be refined, e.g. on a half-day basis.

CONCLUSIONS

This chapter is used less for the technical conclusion of the models than for reflection on the learning process. Accordingly, the argumentation is written in the subjective perspective of the author.

The time taken to obtain the data and analyze them was impressive. I would say that about 70-80% of the working time was used for these two process steps. The previous discussion with the business and in particular with Valiant's data center was time-consuming. Nevertheless, I don't see this time as lost, as the automated data collection and implementation in particular was discussed in these discussions. Furthermore, the selection of models and evaluation metrics was not easy. At the beginning of the work, I wanted to apply all methods in all combinations with all possible parameterization possibilities as comprehensively as possible. Relatively quickly, however, I realized that this effort would be too comprehensive for the given period. Now a compromise was chosen in this work. Next time I would rather focus on even fewer models, but want to delve deeper into matter and analysis. I think that a complete exercise using one model is more meaningful than using as many models as possible, as was done in this project.

The implementation of such a project was very profitable to me; also for Valiant this process was eminently important. This meant that such a project could be conceived within a protected framework and carried out with the necessary departments. The promising forecast means that this project will be continued internally. Internally, another master's thesis with the video call volume will be carried out. This work is supervised by the HSLU. Both models are then compared and applied to the other channel. Thus two neutral perspectives come together and should stimulate each other in order to set up a final model.

REFERENCES

Burger, Scott, Introduction to Machine Learning with R, O'Reilly, Sebastopol, 2018

Hastie, Trevor et al., The Elements of Statistical Learning, Data Mining, Inference, and Prediction, Springer, New York, 2nd Edition, 2017

James, Gareth et al., An Introduction to Statistical Learning with Applications in R, Springer, New York, 8. Edition, 2017

Kuhn, Max und Johnson, Kjell, Applied Predictive Modeling, Springer, New York, 2013

Tables

Figures