About the currents about the development of a model explaining the reasons for the change in the number of contracts for the opening of settlement and cash services in Sberbank, for the purposes of Marketing

June,2020 g.

contents

[1.General Provisions 2](#_Toc12463394)

[1.1. Purpose of Document 2](#_Toc12463395)

[Document 2](#_Toc12463396)

[2.Model 2](#_Toc12463397)  description

[2.1.General description of Model 2](#_Toc12463398)

[2.2.Model 5](#_Toc12463399)  Development Goals

[3.Modeling Data 6](#_Toc12463400)

[3.1.General Description of Data 6](#_Toc12463401)

[4.Building Model 8](#_Toc12463402)

[5.Implementation and limitations of Model 9](#_Toc12463403)

[5.1.Technical implementation of Model 9](#_Toc12463404)

[5.2.Basic Limitations of Model 10](#_Toc12463405)

[6.Модель 11](#_Toc12463406)

[6.1. Feature engineering.](#_Toc12463407)  [11](#_Toc12463407)

[6.2.List of factors. 13](#_Toc12463408)

[6.3.Table of optimized](#_Toc12463409)   [hyperparameters used to convert variables.](#_Toc12463409)  [14](#_Toc12463409)

[6.4. Rating table.](#_Toc12463410)   [14](#_Toc12463410)

[6.5.Exogenousness of variable models. 15](#_Toc12463411)

[6.6.Forecast accuracy score: 16](#_Toc12463412)

[6.7.Checking the model for multicollination. 16](#_Toc12463413)

[6.8.Приложение. 16](#_Toc12463414)

# **General**

## The purpose of the document

This OTchet is designed to describethe modelexplaining the reasons for the change in the number of contracts for the opening of settlement and cash services (RCO) in Sberbank (further model) and ensuring a correct interpretation of the model in the Sberbank units.

To achieve these goals, the Report describes methodological approaches to modeling, as well as the prerequisites and decisions that were made at each stage of the model development.

## Document structure

This report has the following structure:

* Model Description describes the overall logic of the Model calculation
* TheData for Modelingsection describes thesampling procedure for building a model, as well as approaches to data cleanup and processing incorrect data
* The Model Building sectionis dedicatedto describing the general principles of building a model
* The "Model Implementation and Limitations" section describes the limitations of the use of model results.
* The Model section describes the specification and parameters of the model.

# **Description modeland**

## A general description of the modeland

* + 1. The object of modeling (target variable) is to change the number of contracts concluded for the opening of settlement and cash services,as well as to assess the impact of Sberbank's media marketing activities on the target variable.
    2. Coppera-activityis supposed to affect thetarget variable non-linearly.
       - **Advertising**AdStock: takes into account the time-distributed advertising effect; is a type of model with distributedandlagmand.

The adStock effect shows how much % of the advertising effect in the previous period has gone to the current (and so for each measurement):

where at-modeled adfrom advertising in the current period,

Xt isthe level of advertising activity in the current period,

At-1 is a simulated return on advertising in the previous period,

a - the share of the advertising effect flowing from the last period.

* + - * **The saturation effect:** themajesty of advertising increases thepercentage of audience that is achieved through advertising, and therefore increases demand, but the linear increase in advertising impact does not have a similar linear impact on demand.

This transformation is described by the following formula:

????? (2) ,

Where is

?????????????

.

* + 1. The simulation takes place in two stages.
       - The first step is to evaluate hyperparameters describing the effects of saturation and AdStock, using the non-linearmethod of the smallest squares.
       - The second step is based on a linear model on the satiated and AdStock variables that have been transformed with the effect of saturation.
    2. The data is collected daily granularity for the period from 30/10/2017 to 31/03/2020 inclusive.
    3. The data is given to a week-long granularity to eliminate noise during modeling.
    4. The target variable from seasonality was cleaned by STL for the data period 30/10/2017 -1 6/02/2020 20.
    5. The simulation is done on the period from 30/10/2017 to 29/03/2020. This period is divided into two parts: the training period ( 30/10/20/1 6/02/20 20)and the test(17/02/2020/03/2020).
    6. The data is broken down into training and doughsamples in continuous and non-cross dates intervals. Test sample of the model falls on dates from 05/08/2019 to 30/09/2019, training - from 04/09/2017 to 04/08/2019.
    7. The target variable from seasonality has been cleaned up by STL on the training period of the data.
    8. The regression model is chosen because it is not a BlackBox-type model.
    9. Models were used to assess the quality of the models:
       - The following prerequisites are feasible:
         * The remnants have zero mate.
         * No autocorrulation of residues
         * normal distribution of residues
         * No correlation of residues with target variable
       - The stationaryness of the target variable
       - lack of multi-collection of regressors
       - model describes much of the variance of the dependent variable
       - predictive values on the test sample describe a significant portion of the variance of the dependent variable

## Whole model developmentsand

The need to create a model explaining the reasons for the change in the number of contracts for the opening of RCO in the Bank is determined by the objectives:

* Assessing the impact of the Bank's marketing activities
* оценка Media Return-on-Investment (mROI);
* Determining the optimal allocation of the budget to media channels ofcommunication.

Themodel takes into account non-linear effects (such as reduced economies of scale, so-called "saturation curves") and timedistribution (so-called "memory effect") of the effect of marketing activities.

The result of the model will be used for the followingpurposes:

* separating the marketing effect between communication channels (TV, context, internet media,etc.)
* optimizing the marketinof the go-mix between these channels;
* as supportivematerial in negotiations on feasibilityand thegeneral media budget for the RCO product;
* Retrospective calculation of mROI of past advertising campaigns;
* Forecast calculation of mROI of planned advertising campaigns.

# **Simulation data**

## Summary

* + 1. Data was provided by Sberbank:
       - The number of contracts signed for the opening of settlement and cash services.
       - Number of calls from the Bank with the offer of RCO products. The data is fromthe Affiliate Channel division.
       - The number of visitsto the Sberbank siteas a result of contextual advertising in the day-to-day granulation. The data is obtained from advertising offices (in the number of visits). The data is provided by product (RCO, Credit, Business Card).
       - The number of coverages as a result of SMMactivity. Data provided by products (RCO, Credit, Business Card, Now I Boss), by sources(SMM,Bloggers, Crops), by sites(fb, ig, vk, ok, tg).
       - These transitions as a result of online advertising to the pages of blogs "Just about business" and "Own Business."
       - Rates on lending to business by the Bank and competitors in monthly granulation. Information on the average rates of competitors is collected on the basis of publications of transactions on tender platforms (for large-medium business and regional public sector) and on the basis of data collection by RBP-range by a secret buyer (for micro, small, micro-small business).
       - Motivational card of client managers and consultants in GSP on the Bank's products in monthly granulation, as well as a table of performance standards of MP, KBP in monthly granulation.
       - Weighted average interest rates of credit organizations on credit and deposit transactions in rubles excluding Sberbank (% per annum) in monthly granulation.
       - Launch dates for business packages for 0 rubles by the main competitors
       - Number of FNS registrations
    2. Data were collected from the agency:
       - Advertising offline activity of the Bank in the breakdown of the Bank's products (main product - RCO, related products - Credits, credit cards, debit cards) and media channels (television, outdoor advertising, radio, radio regional, press). The data was bought from Mediascope,an industry research agency. Data is provided in thousands of OTS units . OTS: the number of times an audience has been able to contact a particular type of ad.
       - Advertising offline activity of the Bank on special projects
       - Advertising offline activity of the Bank's competitors by RKO product by product by media channels (television, outdoor advertising, radio, radio regional, press). The data was bought from Mediascope. Data is provided in thousands of OTS units .
       - Number of banner displays and OLV on RCO product and related products (Credits, Business Cards, Now I'm Boss).
       - Budgets for digital activity of competitors on the product of RKO
       - Data on offline viewings of the show "Now I'm the boss" on TV. The show includes the Bank's sponsored advertising integrations. Data provided in thousands of OTSunits.
       - Watching "Now I'm The Boss" on YouTube.
       - Number of Yandex requests for RCO products for the Bank and competitors
       - Production calendar.

Table 1 provides the variables that were included in the model and their brief description.

Table 1. **Description of variables of themodel.**

|  |  |  |
| --- | --- | --- |
| **type** | **name** | **description** |
| Thevariable | new\_accounts | Number of contracts signed for the opening of the RCO |
| Seasonal components and non-media factors | d\_day\_type | Number of non-working days per week |
| w\_seas\_y\_clear | The seasonal component of the target variable, cleared from the effects of weekends and holidays. |
| Macro\_nreg\_lag1\_diff | Number of FNS registrations for the previous month |
| d\_quart\_2, d\_quart\_3, d\_quart\_4 | Category variable for 2.3 and 4 quarters |
| Media activity | TV\_TA\_OLV\_Banners\_old | Total OTS (pertarget audience) for TV on RCO products, Credits, Credit Cards, Debit Cards; total OLV and RCO product banners, Loans, BusinessCards. Period - up to 02. 09.2019 |
| TV\_TA\_OLV\_Banners\_new | Total OTS (targetaudience) for TV on RCO products, Credits, Credit Cards, Debit Cards; total OLV and RCO product banners, Loans, Business Cards. 09.2019 |
| calls\_ youwant | The number of calls from the Bank's partners and with the offerofRKO rodukts. The variable is cleaned from the sesamecomponent by STL |
| context\_nocredit\_net\_src\_notv | Total clicks on an ad toa non-textad.  Product: RKO, Business Cards  channels: Search, Networks,Remarketing.  Variable cleared of the impact of TV (OTS on target audience) by the method of the smallest squares |
| context\_credit\_ns19\_notv | Total clicks on an ad toa non-text ad.  Product: Credits;  channels: Search, Networks,Remarketing;  neriod: from 01.01.2019.  Variable cleared of the impact of TV (OTS on target audience) by the method of the smallest squares |
| smm.sb | Total coverage as a result of SMMactivity.  Product: RCO, Business Cards, Credits;  источники: SMM, Блогеры;  плояадки: vk, fb, ig |
| Boss\_TVTA\_YT | Total OTS (target audience) show "Now I'm the Boss!" on TV and on YouTube |
| Offline activity of competitors | Comp2sum\_TV\_comb | Variable, designed to assess the impact of TV-the activity ofcompetitors (seemeaningful interpretation under the table). Total TV OTS (by an audience of 18 degrees) top-11 rival banks. |
| Comp2sum\_TV\_new | Total TV OTS (by an audience of 18q) top 11 rival banks. |

**A meaningful interpretation of some of the variables used for modeling:**

* The variable Comp2sum\_TV\_comb designed to assess the impact of competitors' TV activity on the target variable. The process of building this variable takes place in several stages:

1. TV activity (OTS per target audience) of competitors is based separately on two periods: 30.10.2017-30.06.2019 (Comp2sum\_TV\_old) and 01.07.2019-23. 03.2020 (Comp2sum\_TV\_new). This split is due to the fact that inthe second halfof 2019 competitors began to actively enter the market with new offers for business for 0 rubles (similartothe small start package from Sberbank). In accordance with the changed situation, competitors in these two periods have different influences on the target variable.
2. Each of these two parts has been converted according to the conversion functions, the parameters of which are listed in the conversion table (see paragraph 6.3). Differences intransformations reflect different effects from competitors' activity during relevant periods

The two-part competition is combined for the following reasons:

1. TV activity of competitors on the new period is at the same time as the Bank's TV activity on the new period and has a similardynamic, so the division of competitors into two periods generates multi-coloredness

The weight factor was chosen in accordance with the following circumstances:

1. Preservation of thedefament (key condition).
2. Ensuringoptimal orthogonalization while reducingand multi-collection and maintaining a low forecast error in the training and test period

* Contextual advertising in the model inthe chis in two variables: total clicks on non-credit products (RCO, Business Cards) for the entire period, clicks on credit productam for the period from the beginningof 2019. This division is linked to a disproportionate increase in the volume of amyamof credit products in 2019.
* The model takes into account seasonal changes in target related to weekends and holidays, as well as quarterly seasonality
* Variable TV TA\_OLV\_Banners is also divided into two periods\_OLV\_Banners\_\_OLV\_Banners\_. This difference in effects is displayed in different transformation settings for the respective variables (see Table 2).

# **Building a model**

* 1. Modelis implemented well inthe programming language R.
  2. The model is a logical extension of the basic linear model, but has the following advantages:
     1. The model takes into account the effects of saturation and AdStock media variables, thus more accurately assessing their effects.
     2. The model does not use the target variable as regressors, thereby increasing predictive properties.
     3. Weekends and holidays are not excluded from the data; the model takes into account the decrease in the number of applications for weeks that include holidays.
  3. Two tools are used as a modeling tool:
     1. To select the optimal hyperparameter curves and AdStock use the ?????????? (description of the response curve algorithm)
     2. Linear regression on converted variables after AdStock.

# **Implementation and restrictions**

## The technical implementationof the

Implementation of the model in R-studio software in R allows you to quickly modify the model, supplement the parameters and recalibrate.

The script downloads all the necessary data (pre-processed in advance), builds linear regression, displays model quality graphics into files with .jpeg extension, displays the summary by model in files with the extension of .txt.

Packages used:

* + - * data.table
      * ggplot2
      * tseries
      * car
      * BSDA

To start, you need:

1. распаковать архив rko\_model\_v. 3. 1. zip. Архив содержит следующие файлы:
   * A copy of this report: "The RCO v. 3. 1. docx»
   * File withdata:"rkoqtransformedqdata. Rdata(data contains model variables after conversion.
   * Скрипт модели: «rko\_model\_v.3.1. R»
2. In Rstudio go to the catalog, which is the contents of the archive.
3. Запустить скрипт model\_rko\_v.3. 1. R.

## Key model limitations

* + 1. The model evaluates non-linear processes.
       - Contributions from marketing campaigns are time-limited and accumulate. This is known in literature as the AdStock effect (see. For example. Broadbent 1979,”One Way TV Advertisements Work", Journal of the Market Research Society Vol. 23 no.3.). This is factored into the model.
       - Contributions from marketing campaigns are not linear, and are often modeled in the paradigm ofreducing economiesof scale, so-called saturation curves. Cain 2010, “Limitations of conventional marketing mix modelling.”). This is factored into themodel.
    2. The model does not explicitly take into account political risks and global shock changes in the domestic and external economicsituation.
    3. The model is sensitive to input prediction errors, as well as errors in the actual values of macroeconomic factors.
    4. When extrapolated (using a model on data that goes beyond the values observed in history), the accuracy of the model may decrease because the model was not trained on such data and could not detect a change in functional dependence.

# **model**

## Feature engineering.

In the design of the model, the process of selecting relevant regressors was carried out, taking into account the theoretical foundations set out in the literature on modelingmarketing go mix (see, nexample, Hanssens etal,2001; Ataman etal, 2010; Cain, 2005, 2010; Lodish etal, 2005). Marketing media channels can influence the consumer's desire to purchase a product, increase brand awareness, change the consumer's subjective perception of the price and quality of the product, inform about the specific offer.

Here are some of the key approaches you used to select variables for the model:

* **Media factors. One of the main objectives of the** model is to assess the impact of the Bank's marketing activities on the number of contracts concluded on the opening of the settlement and cash service.
  + Mjedia
    - The numberof spm TV (measuredin 1000 OTS), display advertising and OLV (measured ein1000 views)hasa positive statistically significant effect.
    - other offline media (radio advertising, press advertising and outdoor advertising) are excluded from the model because their contribution is statistically insignificant. This is due to their low number of observations during the period under review.
    - The number of calls from the Bank offering RCO products has a positive statisteechi significant effect on the target variable.
    - statistically significant contribution to the contributionandthat contextual advertising (measuredinthe number of clicks on the contextual ad). At the same time, contextual advertising on RCO products, Basen cards has a greater effect on the target than contextual advertising on loans for business.
    - SMM-promotion (measured in the number of coverages) makesa statistically significant contribution to the target variable
    - The number of views of the show "Now I'm Boss!" on TV and Youtube channel has a positive statstishessignificant effect on the target variable.
  + Offline media competitors
    - Based on the assumption that the advertising of competitors can redirect some potential customers from Sberbank to competitors, the model included offline media **activity**  **top-11**  **largest competitors** (measured in 1000 OTS). The effect of advertising these conkuents in the model is negative. From the second half of 2019, the negative influence of competitors increased due to the launch of competitive packages of services for 0 rubles.

**Exogenousity**. We assume the exogenousness of media factors on the basis of cause and **effect:** media campaigns are planned substantially before the actual implementation of the additional demand for the product arising as a result of these campaigns.

* **Non-media factors**
  + **The dynamics of the number of REGISTRATIONs of the Federal Tax Service** has a positive impact on the target variable.

Below are all the variables and all the conversions that have been tried (only a few have entered the final specification):

* + - * **Seasonality of demand:**
  + Weekly seasonality of opening RCO accounts, allocated with the help of the algorithmA STL for 2.5 years.
  + Weekly seasonality of opening RCO accounts, allocated with the help of the algorithmA MSTL for 2.5 years.
  + Monthly seasonality of opening RCO accounts, allocated with the help ofthe algorithma STL for 2.5per year.
* **The bank's media activity:**
  + Separate media activity via offline channels
  + Total media activity through offline channels
  + Separate SMM-activityby products, sources, sites
  + Total SMM-activity invarious combinations by products, sources, sites
  + Activity in blogs "Own Business" and "Just about Business"
  + Yandex requests of RCO products for the Bank and competitors separately in the original form, separately with clearance from the seasonal component, as well as in the relative indicator of the number of requests of the Bank to the number of requests of competitors
  + Dammy-variables for launch dates by competitors of products for 0 rubles
    - In its original form
    - in the form of their transformations with the help of distributionI Weibullah to take into account the fading effect over time, which were sumted into a single variable
    - in the form of their accumulated sums
    - in theform ofa p.e.v.,consisting of dam-variables for the launch dates of competitors products for 0 rubles, weighted on the share of advertising costs (option 1) and on the share ofthe market (option 2)
* **Media offline activity of competitors:**
  + Total media activity top-11 banks
  + Separate media activity TOP-11 banks (by bank and period)

## A list of factors.

The basic specification is a linear MNC model. The final factors included in the model are listed earlier in Table 1. For orthogonalization, variable ?????????? .

## Table of optimized hyperparameters used to convert variables.

AdStock'Saturation: a combination of features (1) and (2) with specified parameters.

Table 2. **Variable conversion function options.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Transformation function** | **Parameter** | **Value** |
| context\_nocredit\_net\_src\_notv |  |  |  |
| context\_nocredit\_net\_src\_notv |  |  |  |
| context\_nocredit\_net\_src\_notv |  |  |  |
| context\_nocredit\_net\_src\_notv |  |  |  |
| context\_credit\_ns19\_notv |  |  |  |
| context\_credit\_ns19\_notv |  |  |  |
| context\_credit\_ns19\_notv |  |  |  |
| context\_credit\_ns19\_notv |  |  |  |
| TV\_TA\_OLV\_Banners\_old |  |  |  |
| TV\_TA\_OLV\_Banners\_old |  |  |  |
| TV\_TA\_OLV\_Banners\_old |  |  |  |
| TV\_TA\_OLV\_Banners\_old |  |  |  |
| TV\_TA\_OLV\_Banners\_new |  |  |  |
| TV\_TA\_OLV\_Banners\_new |  |  |  |
| TV\_TA\_OLV\_Banners\_new |  |  |  |
| TV\_TA\_OLV\_Banners\_new |  |  |  |
| Boss\_TVTA\_YT |  |  |  |
| Boss\_TVTA\_YT |  |  |  |
| Boss\_TVTA\_YT |  |  |  |
| Boss\_TVTA\_YT |  |  |  |
| smm.sb |  |  |  |
| smm.sb |  |  |  |
| smm.sb |  |  |  |
| smm.sb |  |  |  |
| Comp2sum\_TV\_old |  |  |  |
| Comp2sum\_TV\_old |  |  |  |
| Comp2sum\_TV\_old |  |  |  |
| Comp2sum\_TV\_old |  |  |  |
| Comp2sum\_TV\_new |  |  |  |
| Comp2sum\_TV\_new |  |  |  |
| Comp2sum\_TV\_new |  |  |  |
| Comp2sum\_TV\_new |  |  |  |

## Score table.

Warning: Simulation results and statistical tests are stored in the"rkov"file. 3. 1-model-tests. txt" in the work folder.  10.2017-16.02.2020,Testrange: 17.02.20 20 - 29.03.20 2020.

Adjusted R2 = 0.9695

Box-Ljung test (остатки): X-squared = 1.1729, df = 1, p-value = 0.2788

Augmented Dickey-Fuller test: Dickey-Fuller = -5.1104, Lag order = 4, p-value = 0.01

The match test is amate. expectations of residues on the validation set are zero: the forecasts are not shifted.

z = -0.72194, p-value = 0.4703

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | east | SD | t-statistics | P>t | Significance |
| (Intercept) |  |  |  |  | \*\*\* |
| context\_nocredit\_net\_src\_notv |  |  |  |  | \*\* |
| context\_credit\_ns19\_notv |  |  |  |  | \*\*\* |
| TV\_TA\_OLV\_Banners\_new |  |  |  |  | \* |
| TV\_TA\_OLV\_Banners\_old |  |  |  |  | \*\*\* |
| Boss\_TVTA\_YT |  |  |  |  | \* |
| smm.sb |  |  |  |  | \* |
| Macro\_nreg\_lag1 |  |  |  |  |  |
| d\_quart\_2 |  |  |  |  | \* |
| d\_quart\_3 |  |  |  |  | \*\* |
| d\_quart\_4 |  |  |  |  | \* |
| d\_day\_type |  |  |  |  | \*\*\* |
| w\_seas\_y\_clear |  |  |  |  | \*\*\* |
| PC1 |  |  |  |  | \*\* |
| PC2 |  |  |  |  | \*\* |

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

## The exogenity of variable models.

* Media variables: the exogenity of media variables is caused by causal logic. Namely, the decision on the volume of advertising is made significantly before its implementation.
* Thus, the regressors are exogenous, and linear regression is used to improve the model's effectiveness.

## Assessment of the accuracy of the forecast:

На train-sample (30.10.2017-16.02.2020) MAPE равен 5.31, на test-sample (17.02.2020 – 29.03.2020) MAPE равен 2.39.

## Checking the model for multicollinarity.

The VIF (variance inflation factor) criterion was used to test the model for multicollinarity. As you can see from the table, there is no significant multi-collectionivity in the model(80% of VIF-factorvaluesare less than 3 and 20% of VIF-factor values are less than 5).

|  |  |
| --- | --- |
| **Variable** | **bright** |
| context\_nocredit\_net\_src\_notv | 2.359102445 |
| context\_credit\_ns19\_notv | 2.263367017 |
| TV\_TA\_OLV\_Banners\_new | 4.063089188 |
| TV\_TA\_OLV\_Banners\_old | 2.374878411 |
| Boss\_TVTA\_YT | 1.821826658 |
| smm.sb | 1.685171611 |
| Macro\_nreg\_lag1 | 1.629053458 |
| d\_quart\_2 | 1.937875103 |
| d\_quart\_3 | 2.611728446 |
| d\_quart\_4 | 2.267826273 |
| d\_day\_type | 2.505774283 |
| w\_seas\_y\_clear | 1.450454339 |
| PC1 | 4.939332749 |
| PC2 | 3.597259373 |