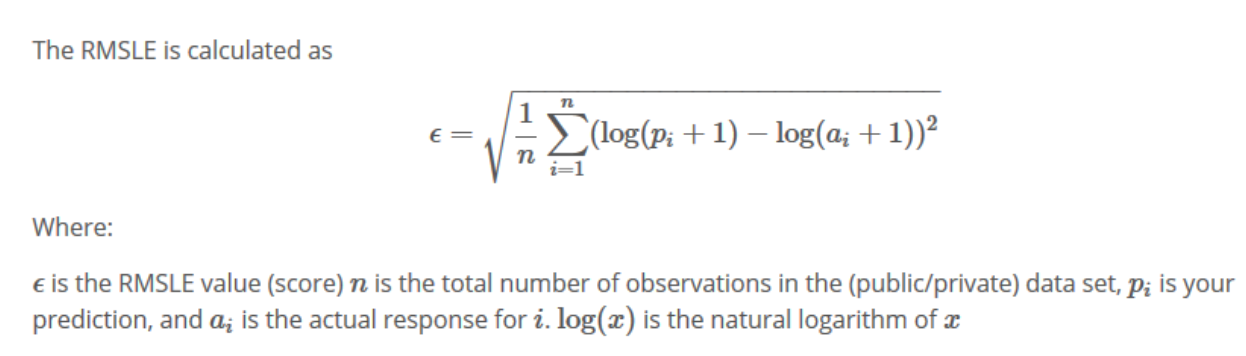


Housing costs demand a significant investment from both consumers and developers. And when it comes to planning a budget—whether personal or corporate—the last thing anyone needs is uncertainty about one of their biggest expenses. Sberbank, Russia’s oldest and largest bank, helps their customers by making predictions about realty prices so renters, developers, and lenders are more confident when they sign a lease or purchase a building.

Although the housing market is relatively stable in Russia, complex interactions between housing features such as number of bedrooms and location are enough to make pricing predictions complicated.

In this project, you are challenged to develop models which use a broad spectrum of features to predict realty prices. An accurate prediction model will allow Sberbank to provide more certainty to their customers and value to their shareholders. There are two goals:

1. We would like to build a useful model to predict the price of an individual property given some of the variables in the data dictionary and in the dataset *modelingData.csv*. You will make predictions using the model you develop using *predictionData.csv* which contains 7662 properties that sold between July 2015 and May 2016. Your goal is to derive a model that minimizes the RMSLE (Root Mean Squared Logarithmic Error) as defined below.



1. We would like to build a useful model to forecast the **mean** price of properties from July 2015 to July 2016. We would also like to provide a table and a plot of these forecasts complete with 95% confidence intervals. This analysis will entail extensive data wrangling. In addition, the analysis should test for serial correlation in the data and make necessary adjustments to the model if there is evidence of significant serial correlation.

More details on both of these goals are below:

**Goal 1: Prediction of Individual Property Values 8 Pages (*60 pts*)**

It should take 8 or less pages to address Goal 1. Your analysis should include the following:

1. Introduction (Brief introduction to the problem … 1 or 2 paragraphs).
2. Description of the data with a table or a reference to a table in an Appendix.
3. Data Cleaning / Wrangling (any renaming of variables or standardizing of values.)
4. Exploratory Data Analysis (EDA).

(Complete with summary statistics, descriptions, tables and/or plots etc.)

* 1. Outlier Identification and Handling
  2. Missing value identification, summary and possible imputation (mean, median, regression.) This may also be considered “Data Wrangling”.
  3. Multicollinearity (is there reason to believe it is present?) You don’t have to address every potential pair of variables that may be collinear. Just provide a plot and or other evidence of a single occurrence of multicollinearity if at least one exists and then mention possible other occurrences.
  4. Checking assumptions: Homoscedasticity, normal distributions of the response for fixed values of the explanatory variable(s), linear relationship between the mean of the response and each explanatory variable, etc. This is where you would apply transformations (log, square root, etc.)
  5. Variable selection: For example, there are many potential explanatory variables. Running stepwise variable selection will not necessarily provide a final model, but may leave you with a smaller set of potential explanatory variables to work with.
  6. Anything else that might be appropriate in learning about the data before getting started. (Example: You might analyze interactions between explanatory variables in the analysis.)

\*Note: For a refresher on Exploratory Data Analysis (EDA), review the EDA Unit in Doing Data Science.

1. Modeling
   1. You must fit at least 3 candidate models.
      1. A model with OLS parameter estimates. You may choose the variables with or without the use of a variable selection technique (forward, backward, stepwise) (You may have done this already in your EDA to pare down the number of variables to work with.)
      2. A model with LASSO estimation and selection.
      3. A model of your choice. This may be using another OLS or LASSO model or custom model, etc.

\*Note: In at least one of the models above, or in your EDA, you should test the significance of at least one term that involves an interaction.

* 1. You need to address the assumptions with respect to the residuals. (Normally distributed around 0 with constant standard deviation.)
  2. For each model you need to conduct an internal and external cross validation.
  3. You should compare the models using the AIC, SBC, internal k-fold cross validation with the CV Press Statistic (you pick k), and external cross validation. For external cross validation you will have to subset the data set, *modelingData.csv,* into train and test data sets.

1. Prediction
   1. Identify which model you feel is the best and discuss why.
   2. Use this model to predict the price of each of the 7662 properties in the prediction data set (*predictionData.csv*).
   3. Submit these predictions in the zip file with your final paper. Please see the sample submission file. *SampleSubmission.csv*

\*Note: It is understood that all the material above should be supported with plots, tables, charts, discussions, etc. where appropriate. All plots and tables should be clearly labeled and referenced in the discussion and all tables referenced in the discussion should exist in the paper. It is never appropriate to have a plot or table in the paper that it not described / referenced in the writing. *Your team could lose points or gain points depending on how well your professor feels you described research / findings.*

**Goal 2: Prediction of Mean Property Value by Year 3 Pages (*30 pts*)**

1. Introduction (Brief introduction to the problem … 1 or 2 paragraphs).
2. Data Wrangling: Here you will have to manipulate/wrangle the data to produce the data you will use in this analysis. Your goal is to get the mean price for each month/year combination. Hints:

Data Wrangling Use SAS, Python or R (hint: dplyr and tidyr packages) to separate the timestamp column into 3 columns. These three columns should be called “Day”, “Month” and “Year” respectively (any order is fine). To be clear, your task is to turn:

 into 

USE R,Python or SAS to aggregate the data by year and month. This means that the new dataset should have the mean of all the properties for a given month in a given year. You may do this using any method you like. The “aggregate” function in R is one way of doing this. As a reference, the first four months should match the values below.



1. Plot the time series of the price\_doc versus the numbers 1 – 47 on the x-axis. You should simply name the x-axis “months”.
2. Model the residual series.
   1. Fit a simple linear regression model with price\_doc as the response variable and Month\_Number as the explanatory variable.
   2. Obtain the residuals from this model and plot them against the month number. Make sure this plot is a series rather than a scatter plot. This simply means make sure the points on the plot are connected by a line. Keep the x-axis label: “Month”.
   3. Fit the residual and time data using proc autoreg and investigate the autocorrelation structure based on the Durbin-Watson statistic, partial autocorrelation plots, AIC and SBC. You studied the AR(1) model in depth in Unit 4, it is possible you may find that an AR(p) with p > 1 is a better fit to this data. Explore what this means in terms of intuition, model construction and forecasting.
   4. Use your model from part c to forecast the residual for the next year (June 2015 – June 2016) with 95% confidence intervals.
3. Using your model from the last question and the series of residuals, forecast the next year: July 2015 – June 2016 and of course include 95% confidence intervals. You can do this by obtaining an estimate from the model in part b for the mean price (the trend estimate) and then add to it the forecast residuals from the model in part d. Squaring the forecasted error term gives an estimate of the variance of the predicted price that can be used to calculate confidence intervals. The confidence interval won’t be exact, but a conservative estimate will be to add/subtract the square root of the forecasted squared error term times 1.65 to/from the predicted price to give the upper/lower bounds for each prediction. SAS also provides predictions and confidence intervals.
4. Submit your forecasts (with confidence intervals) as well as your final csv data set you used in the analysis (wrapped in a zip file.)

**Appendix 1: Data Dictionary *3 pts***

1. This Appendix does not count against your page count.
2. You should elaborate on the data dictionary you have received. Add color, add examples, add data types, change the formatting. Do something to improve the look and/or readability!

**Appendix 2: Code *7pts***

(Well commented … Remember: Reproducible Research!)

1. This Appendix does not count against your page count.
2. Simply cut and paste your well commented code in the Appendix.

**Submissions:**

What to submit 2 DS in a single zip file:

1. Prediction from Goal 1. (csv file)
2. Aggregated data set from Goal 2. (48 rows including title row. / csv file)
3. Predictions from Goal 2. (csv file)
4. Line plot of predictions from Goal 2 with 95% confidence intervals. Image or cut and pasted into something like a word doc.
5. Final paper (No longer than 11 pages without appendix.)

**Note: Data Wrangling:**

*Wrangling = having a long and complicated dispute.*

Part of this project is meant to have a significant data wrangling component. As an example, you will more than likely need to work with R or SAS or both to change data from character/string to integer/numeric so your models make the predictions that are required. This is only an example of the data wrangling you will need to conduct. It will help to start early and bring these issues up in live session and/or office hours.

**Due Date:**

**All submissions are due no later than 10:00am CST Saturday February 18th.**