A PROJECT REPORT ON

RUSSIAN HOUSING MARKET

Date: 18 June 2017

PRESENTED BY :	HEMANT KUMAR SAIN
SUBMITTED TO:	EDWISOR.COM

ACKNOWLEDGEMENT

Project development is not an easy task. It requires corporation and help of various people. It always happens that word run out when we are really thankful and sincerely want to inspire my feelings of gratitude towards the one when helped in the completion of the project.

I would like to give my sincere thank to whole Edwisor team, for Giving me an opportunity to learn under them and guided me throughout the project gained comprehensive knowledge of various aspects of how to approach an analytical problem and analyze it. Without their knowledge i would have never been able to complete my project.

I would also like to convey my sincere thank to Mr. Muquayyar Ahmed, Ayush Choudhary and Papori Goswami for their valuable comments and suggestion that has helped in completing my course and Project as well.

A special thanks to all my fellows for helping me out on discussion board with all my queries from the first day of my learning to throughout the project.

I also obliged to whole Edwisor team and all members for their valuable support throughout.

.

Hemant Kumar Sain

CONTENT

Chapter 1: Sberbank Russian Housing Market

1.1 Problem Statement

Chapter 2: Data Preparation

2.1 Data Collection

2.2 Data Cleaning

2.3 Exploratory Data Analysis

Chapter 3: Model Building

3.1 Brief of Recursive Partitioning

3.2 Cross Validation

Chapter 4: Model Deployment

Chapter 5: Major Challenges Faced In this Project

Conclusion

References

CHAPTER-1

SBERBANK RUSSIAN HOUSING MARKET

1.1: Problem Defination:

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 biggets 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, the country’s volatile economy makes forecasting prices as a function of apartment characteristics a unique challenge. Complex interactions between housing features such as number of bedrooms and location are enough to make pricing predictions complicated. Adding an unstable economy to the mix means Sberbank and their customers need more than simple regression models in their arsenal.

In this competition, Sberbank is challenging Kagglers to develop algorithms which use a broad spectrum of features to predict realty prices. Competitors will rely on a rich dataset that includes housing data and macroeconomic patterns. An accurate forecasting model will allow Sberbank to provide more certainty to their customers in an uncertain economy.

CHAPTER-2

DATA PREPARATION

Data preparation (or data preprocessing) in this context means manipulation of data into a form suitable for further analysis and processing. It is a process that involves many different tasks and which cannot be fully automated.

There is few steps explained below which is used to prepare the data for model building.

1. Data Collection
2. Data Cleaning
3. Exploratory Data Analysis

2.1. Data Collection:

Data is given in two sets i.e Test.CSV and Train.CSV

All the operation suppose to perform on both datasets.

2.2. Data Cleaning

To perform this task all the library which to be used in our whole process is initialized and after that basic data cleaning technique is applied to remove the noise in present in data.

Steps taken for data cleaning is mentioned below-

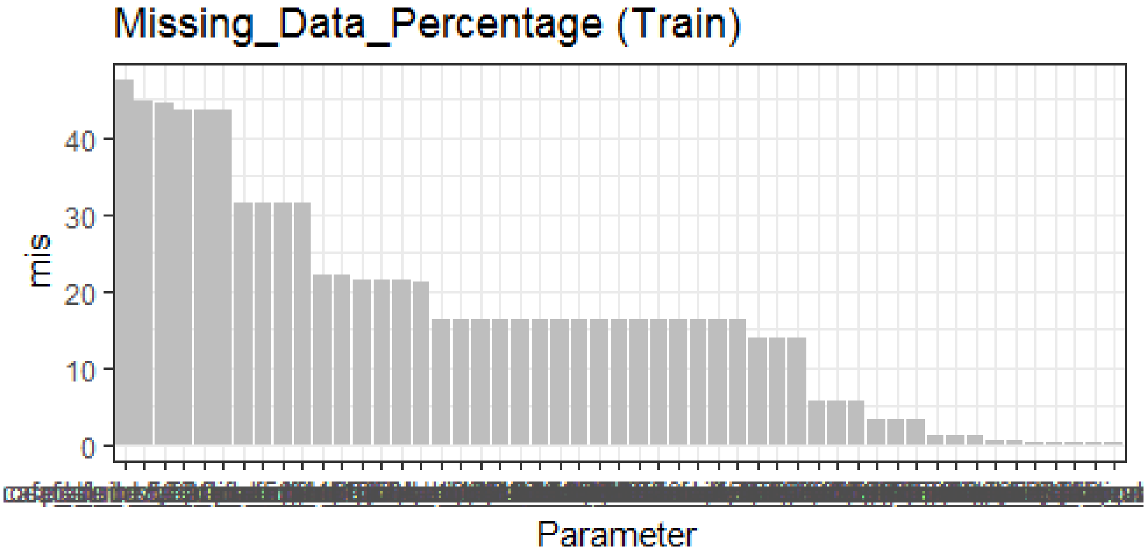
- **Missing Values Detection.**
- **Missing Values Visualization.**
- **Missing Values Imputation.**
- **Outlier Removal.**
- **Data Normalization.**

2.3. Exploratory Data Analysis

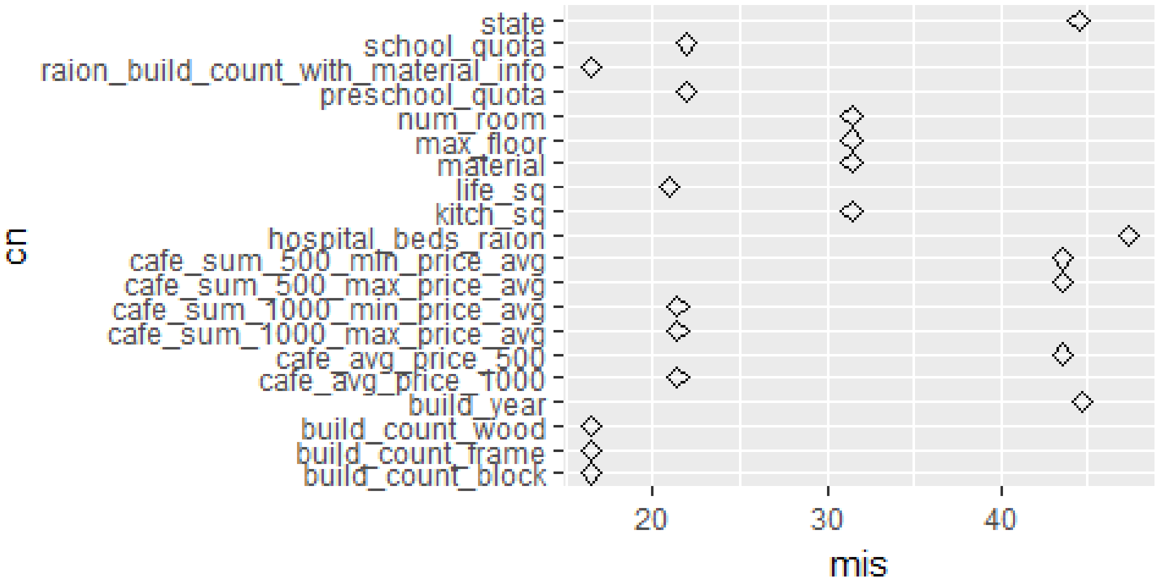
Exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods.

In this case we have taken serval steps like visualizing for various graphs for basic understanding of data.

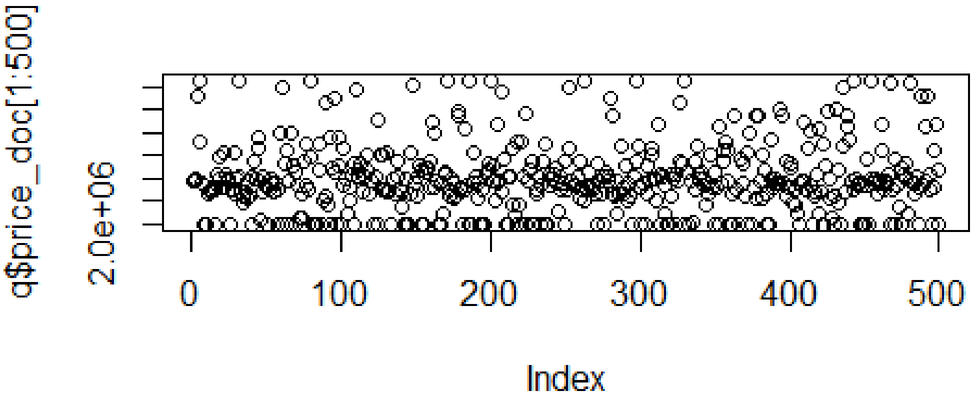
All the steps is explained by step by step under below.



(Fig. Missing Data Percentage vs Variables Bar Plot)

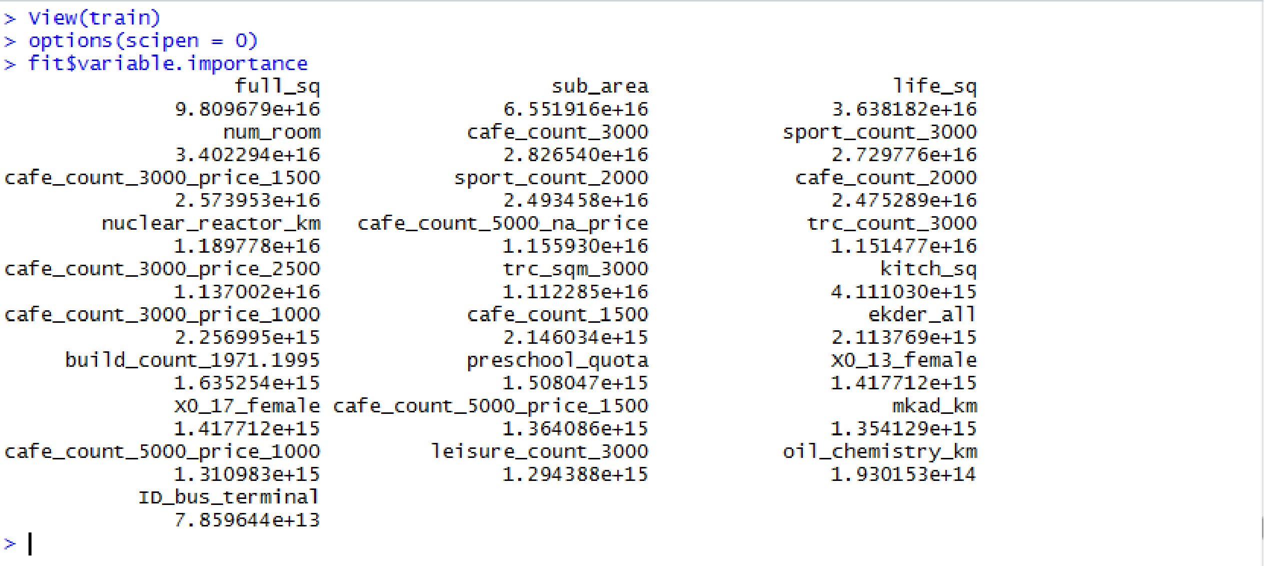


(Fig. Missing Values Percentage VS Variable Scatter Plot)



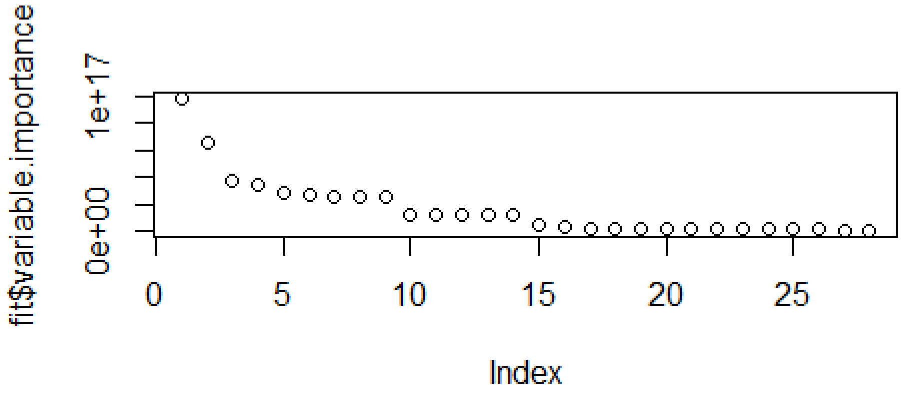
(Distribution of Target Variable)

Variable Selection: We have got high dimensional data having 292 variable which needs to be reduced for better model so we have used varImp function which calculates the importance. And top 17 variables are selected as predictors.



```
> View(train)
> options(scipen = 0)
> fit$variable.importance
                   full_sq                   sub_area                    life_sq
              9.809679e+16               6.551916e+16               3.638182e+16
                  num_room            cafe_count_3000           sport_count_3000
              3.402294e+16               2.826540e+16               2.729776e+16
cafe_count_3000_price_1500           sport_count_2000            cafe_count_2000
              2.573953e+16               2.493458e+16               2.475289e+16
        nuclear_reactor_km   cafe_count_5000_na_price             trc_count_3000
              1.189778e+16               1.155930e+16               1.151477e+16
cafe_count_3000_price_2500               trc_sqm_3000                   kitch_sq
              1.137002e+16               1.112285e+16               4.111030e+15
cafe_count_3000_price_1000            cafe_count_1500                  ekder_all
              2.256995e+15               2.146034e+15               2.113769e+15
     build_count_1971.1995            preschool_quota               X0_13_female
              1.635254e+15               1.508047e+15               1.417712e+15
              X0_17_female cafe_count_5000_price_1500                    mkad_km
              1.417712e+15               1.364086e+15               1.354129e+15
cafe_count_5000_price_1000         leisure_count_3000           oil_chemistry_km
              1.310983e+15               1.294388e+15               1.930153e+14
           ID_bus_terminal
              7.859644e+13
> |
```

.



(Fig. Variable Selection)

CHAPTER-3

MODEL BUILDING

In this phase processed data is used to build model, Regression tree algorithm using CART is used to build regression model.

Preprocessed data has normalized before applying Decision tree algorithm on the top of it so that range of each variable can be equalized.

Internal mechanism of regression tree can be explained as follow

How Decision Tree using rpart works:

rpart algorithm works by splitting the dataset recursively, which means that the subsets that arise from a split are further split until a predetermined termination criterion is reached. At each step, the split is made based on the independent variable that results in the largest possible reduction in heterogeneity of the dependent (predicted) variable.

Splitting rules can be constructed in many different ways, all of which are based on the notion of impurity- a measure of the degree of heterogeneity of the leaf nodes. Put another way, a leaf node that contains a single class is homogeneous and has impurity=0..

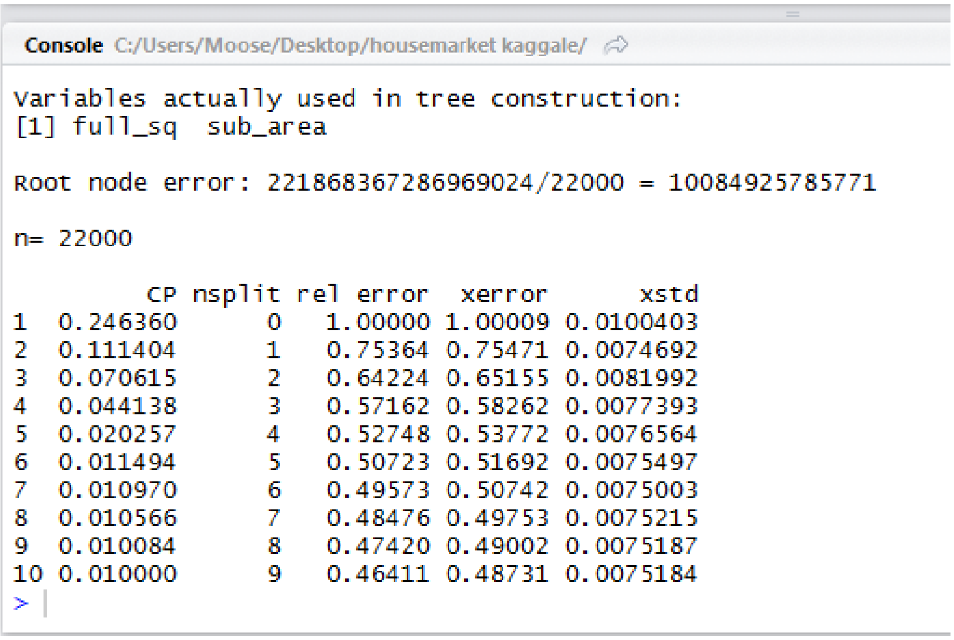
The rpart algorithm offers the entropy and Gini index methods as choices. There is a fair amount of fact and opinion on the Web about which method is better

Let’s now look at the case in which the predicted variable is continuous.

Next we invoke rpart, noting that the predicted variable is **price_doc** and that we need to set the method parameter to “**anova**“. The latter tells rpart that the predicted variable is continuous (i.e that this is a regression problem).

For model building whole train data is again divided into two parts as train and test 70% and 30% respectively of train data.

A quick description image of rpart model is showing under below

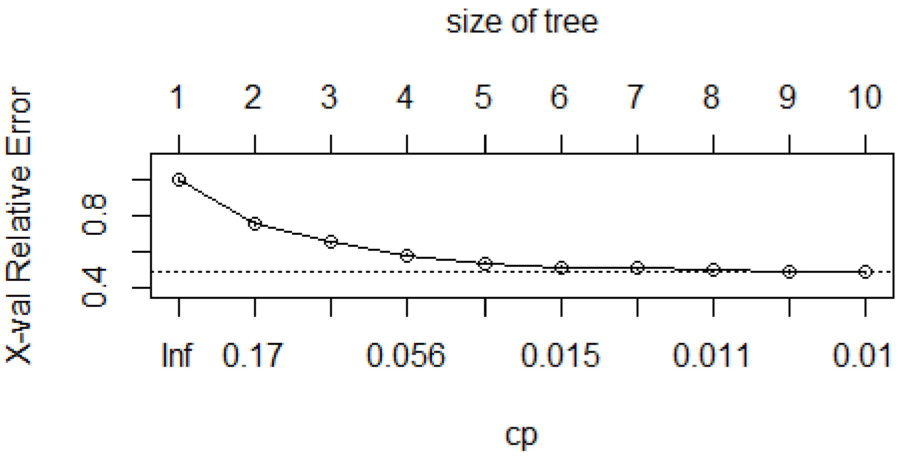


```
Variables actually used in tree construction:
[1] full_sq  sub_area

Root node error: 221868367286969024/22000 = 10084925785771

n= 22000

         CP nsplit rel error  xerror      xstd
1  0.246360      0   1.00000 1.00009 0.0100403
2  0.111404      1   0.75364 0.75471 0.0074692
3  0.070615      2   0.64224 0.65155 0.0081992
4  0.044138      3   0.57162 0.58262 0.0077393
5  0.020257      4   0.52748 0.53772 0.0076564
6  0.011494      5   0.50723 0.51692 0.0075497
7  0.010970      6   0.49573 0.50742 0.0075003
8  0.010566      7   0.48476 0.49753 0.0075215
9  0.010084      8   0.47420 0.49002 0.0075187
10 0.010000      9   0.46411 0.48731 0.0075184
> 
```

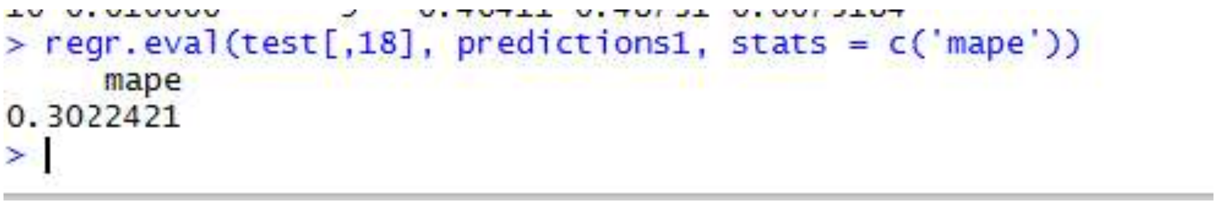


(Fig. Complexity Parameter of Model)

3.2: Cross Validation :

mean absolute percentage error calculated as error matrix which state that how well model is performing it is expressed as a ratio of difference in the estimated value with the actual value divide over every observation.

as we can see value of MAPE is shown 30% in our case means 70% of accuracy serving by our model



```
> regr.eval(test[,18], predictions1, stats = c('mape'))
     mape
0.3022421
> |
```

(Fig. Mean absolute percentage error)

CHAPTER-4

MODEL DEPLOYMENT

In this phase model can be applied on desired dataset and classified values can be predicted.

In this case test data set which is provided having 7662 of observation.

All the preprocessing steps are applied to this test data set as before, and finally applied Previosly Decision tree model on the top of it.

CHAPTER-5

MAJOR CHALLENGES FACED

There was several challenges I have faced during the project, some of measure challenges are listed below -

- Data cleaning and missing value imputation is the most challenging task for this project because of high dimensionality of the data

CONCLUSION

By looking into MAPE, accuracy of model can be considered as 70%, it means this model is suppose to forcaste approximate upto 70% accurate values of test data set with same accuracy.

REFERENCES

- https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf
- https://edwisor.com/myskills.html