Eleven - The right price

Hackathon DSBA - Group 1
Vanille BOURRE, Chih-Tung CHEN, Nan CHEN, Hugo CHIKLI, Jinji SHEN, Adel REMADI

Table of content

Optimize prospection strategy to enhance capital gains

The client, a player in real-estate industry, wants to increase its financial results.

We recommend two levels of improvement on your prospection strategy to increase your returns.

1. Estimate the expected profits for identified locations

Leverage on open data from official mutation databases in Île-de-France.

Build a simulation tool tailored to any real estate project powered by machine learning models.

2. Identify the best locations to consider for future projects

Identify the top 10 areas where the expected profits are the most promising for real-estate project.

Build a simulation tool to identify the areas with highest expected returns.

1. Estimate the expected profits for identified locations

The approach is structured around three complementary axes leading to several deliverables, all embedded in one application

Explore the market

Visualize the state of the real estate market in Ile-de-France and analyse the different market trends according to selected parameters.

Feature engineering based on:
- Time
- Location
- Surface of building
- Number of apartments
- Number of rooms/apartment
- Location

Predict the prices

Use a machine learning model to estimate price based on given criteria.

A real-estate project involving construction and sale takes up to 5 years.

Combining an Ensemble and Time-Series approach, our model predicts expected future prices.

Estimate profits

Account for the cost of real-estate projects to provide expected profit estimate for a given project.

A benchmark has been led to estimate the cost of real-estate construction and renovation projects.

Our model computes expected returns on any given time horizon.

1. Estimate the expected profits for identified locations

The Price Estimation pipeline, combines a tree-based Ensemble algorithm (XGBoost) and a Time-Series analysis:

Approach

- Prepare a baseline model
- Review and implement state-of-art algorithms
- Analyze the relevance of external features
- Select the algorithm based on:
    1. Explainability
    2. Average error
    3. Cross-Validation
- Hyper-parameters tuning

Final Pipeline & Results

XGBoost

The XGBoost predicts the current price value based on selected features (location, surface, nb. of apartments…)

↓

Time-Series

The future expected price value is estimated with several models to capture the trend for each 'department'.

Obtained results on the test set

Absolute Error (MAE)	**35 416.97 €**
Absolute Error Ratio	**13.34 %**

Next Steps

- Further tune the hyper parameters to generate a better zoning (locations).
- Include internal features based on the client's historical data.
- Add relevant external features to reduce variability.
    - Balcony/Terrace
    - Number of bathrooms
    - Isolation rating
    - Consumption rating

2. Identify the best locations to consider for future projects

To gain competitive advantage, it is important to be able identify where to focus the next prospective efforts.

Identify future promising locations

The algorithm screens across the 1200 zip codes in IDF to identify the best prospects in terms of expected profits.

Tailor queries to your projects

The promising areas are identified taking into account your project specifications:

- Surface targeted
- Number of Apartments
- Number of rooms

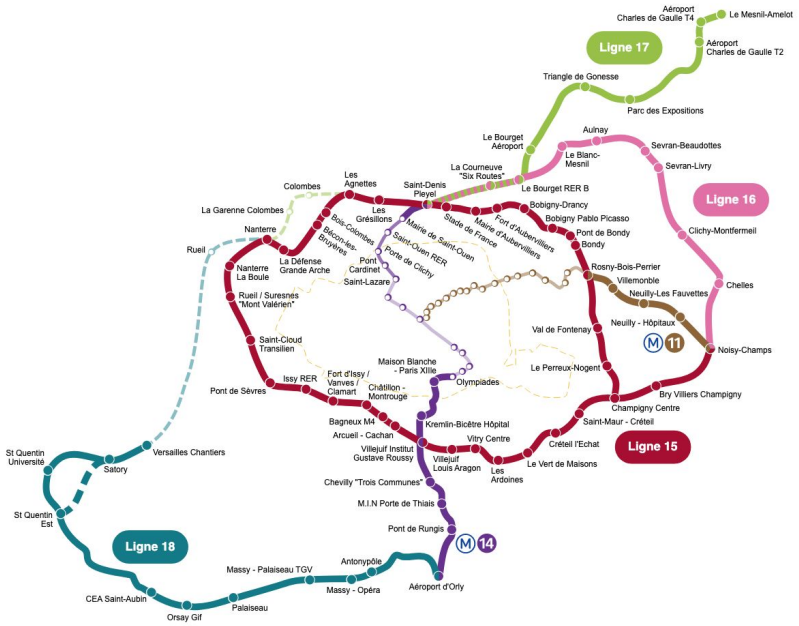
Next steps, Opportunity, Risks

Opportunities & Next step

- Project ‘Grand Paris’ can be a generator of growth
- In the future, a model to optimize the composition of a project given a surface and location could be built:
  How many apartments ? How many rooms ?

Risks and Limitations

- There is still a significant part of the variance not covered by the model.
- Other relevant features might enrich the model.
- It may be a challenge to predict properly the effect of Project ‘Grand Paris’



Simulation tool Demonstration

APPENDIX

ROI Dashboard

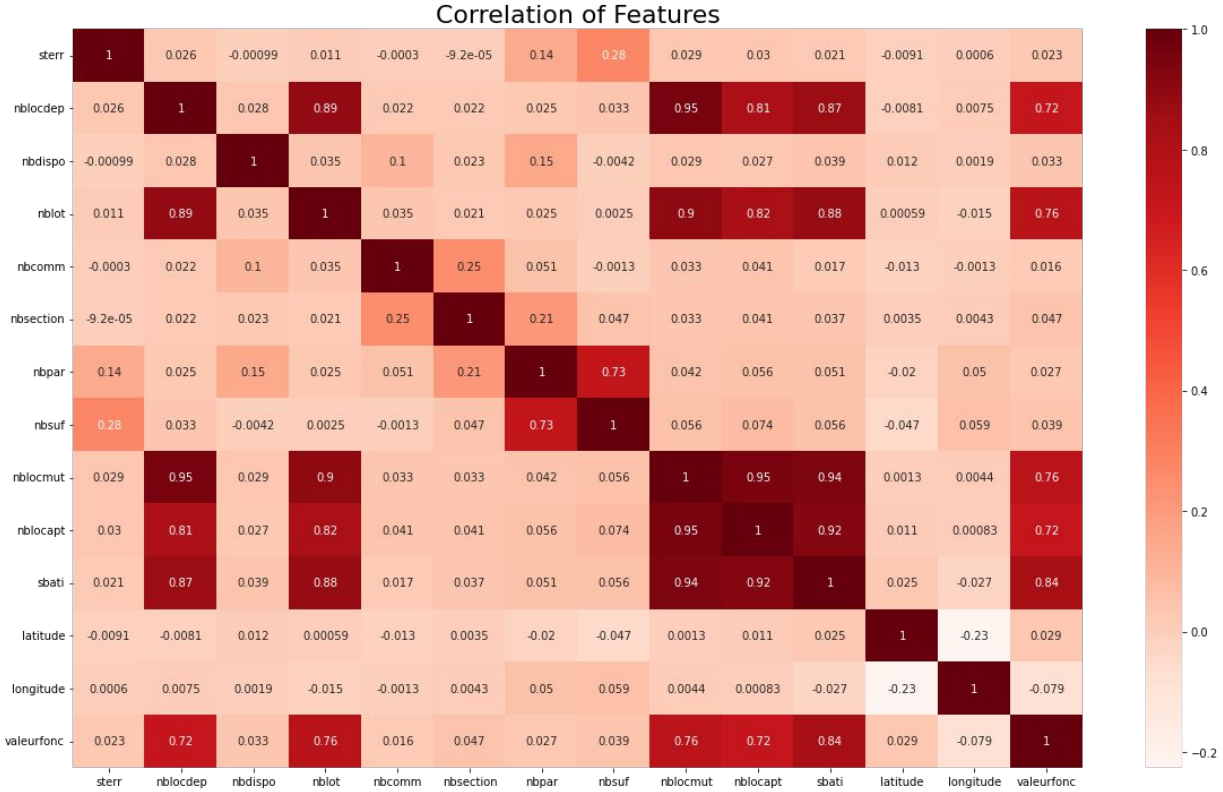
Select Your Parameters
Enter square M^2
2000
Enter Departement Code
75
Construction or Renovation?
construction
horizon
5
Rent or Sell?
Rent

*After 2 years of building

Summary of Return on Investment

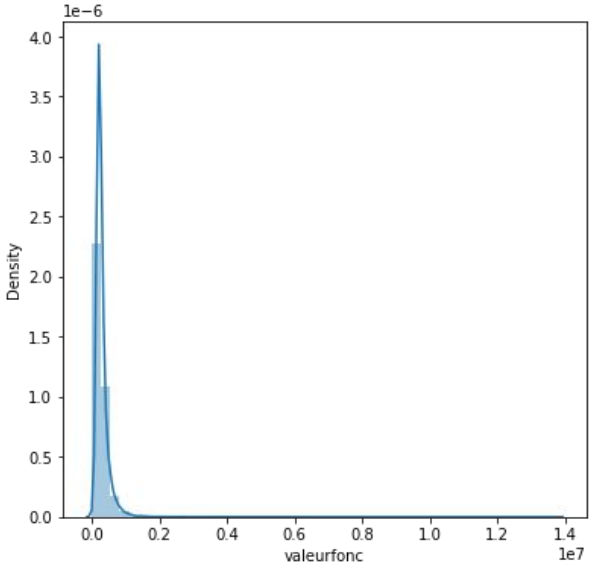
		Over the entire horizon	*Over one year*
Revenue	Rent Revenue*	€2 304 000	€460 800
	Other Revenue		
Cost	Land Cost	-€18 664 093	-€3 732 819
	Building Cost	-€3 500 000	-€700 000
Capital Gain	New Property value	€37 573 387	€7 514 677
	Total Revenue	€2 304 000	€460 800
	Total Cost	-€26 830 117	-€5 366 023
	Capital Gain	**€13 047 270**	€2 609 454
	ROI	**49%**	10%

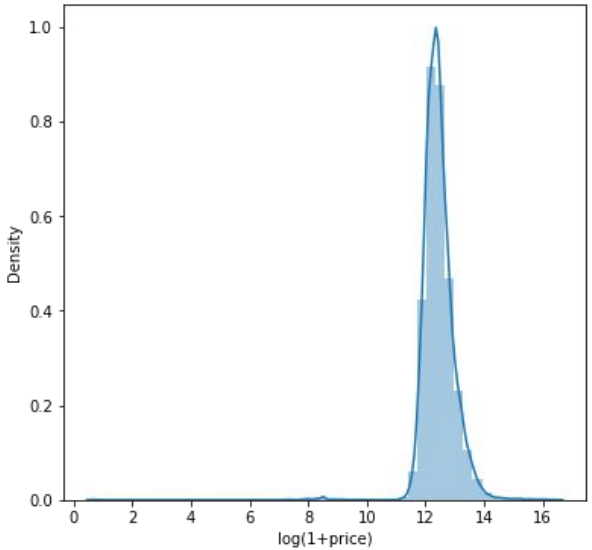
Correlation of features

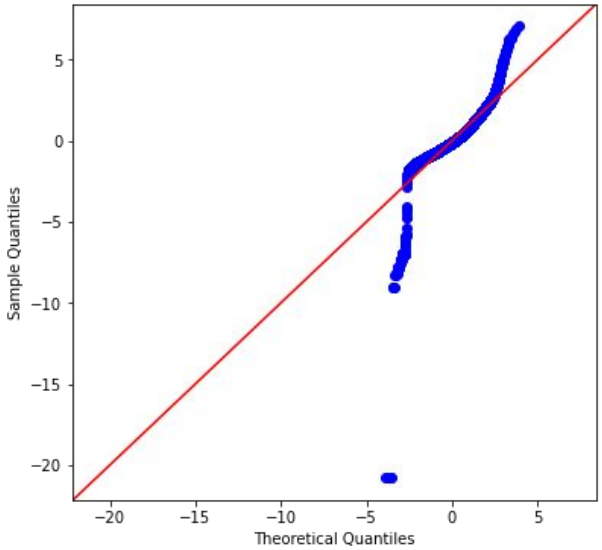


Distribution of New Apartments' Prices

The price distribution is very skewed and long tailed.







Feature Engineering

Data & Algorithm: We tested several machine learning methods to select the best methods and data enhancement techniques from dozens of models

Data collection

data.gouv.fr

Collected past real estate transactions using open data

îledeFrance mobilités

Computed the distance to the nearest subway station

Feature Engineering

- **Time**
  - 'day','month','year'
- **Surface of building**
  - 'sbati'
- **Number of flats & Number of rooms per flats**
  - 'nbapt1pp','nbapt2pp','nbapt3pp','nbapt4pp','nbapt5pp'
- **Location**
  - 'Coddep' one-hot encoded
  - 'Latitude','longitude' **converted in cartesian coord.**

Models tested

Regression:

- Linear Regressions
- Lasso
- Ridge
- Robust (huber)
- Random Forest
- XGboost

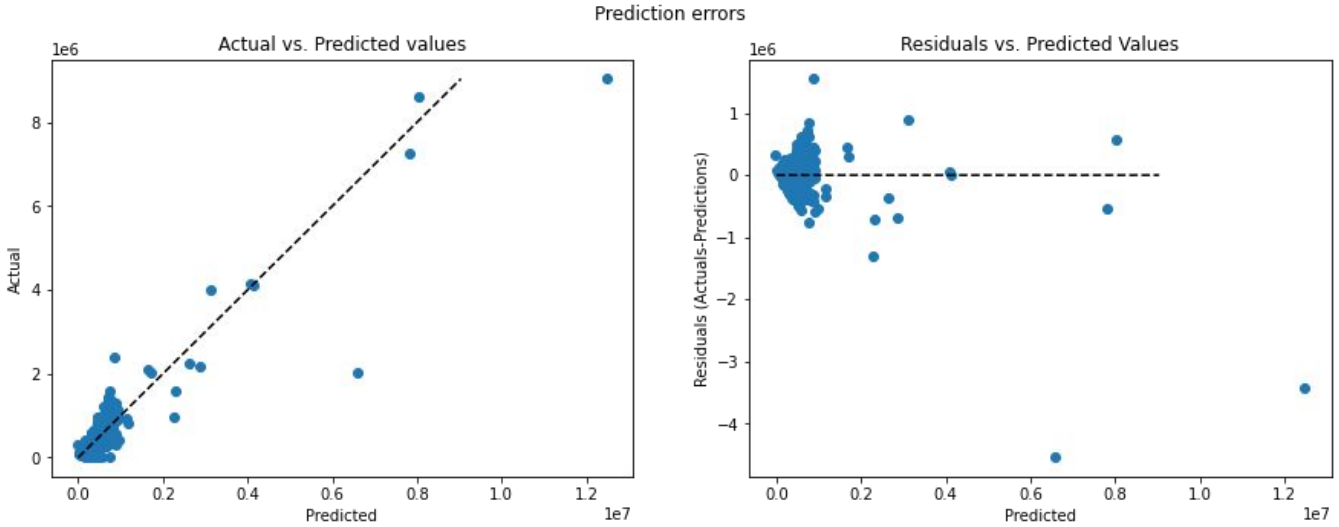
Time series:

- We ran stationarity tests and ACF/PACF analysis.
- We built 1 model per department.
- Monthly average + Linear Regression.

CentraleSupélec

Linear Regression Results on a validation set

A Linear Regression gave baseline scores on a validation set.



Metric	Score
R^2	0.7978
MAE	64080.55
RMSE	137020.37
MAPE	97.644
Average Ratio RMSE	45.87 %
Average Ratio MAE	21.45 %

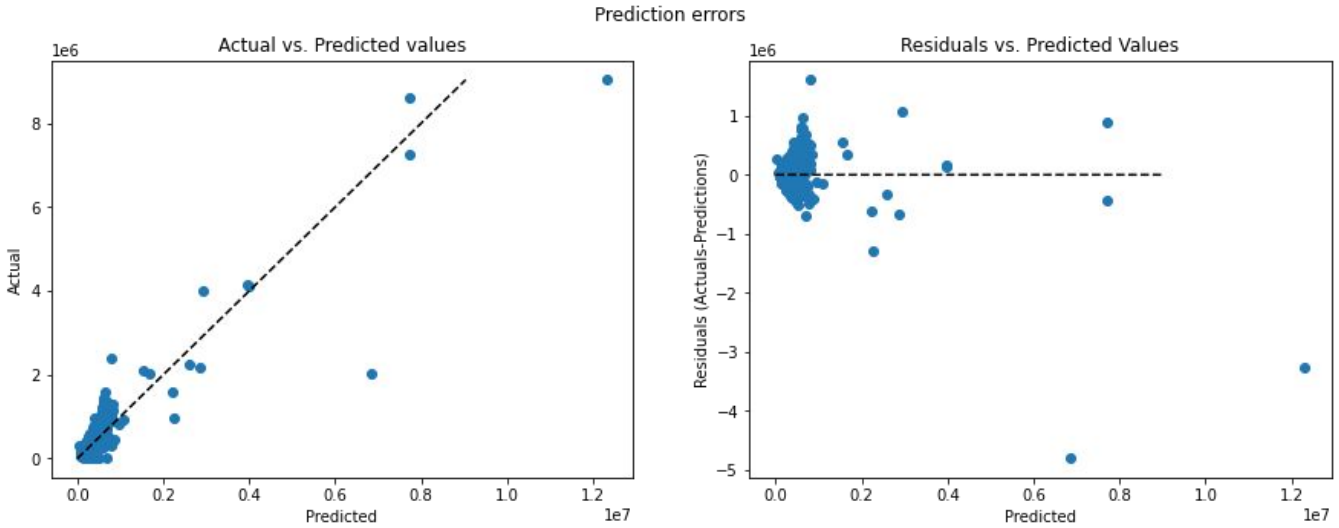
Linear Regression Cross-Validation

Results of Cross-Validation on 5-folds

Fold	R^2	RMSE	MAE	MAPE	Average Ratio RMSE	Average Ratio MAE
1	0,5672	248671,73	72057,74	98,27	0,83	0,24
2	0,6115	159854,97	65526,91	0,43	0,54	0,22
3	0,7046	166330,03	67330,21	69,95	0,56	0,23
4	0,8363	169136,72	67784,87	45,21	0,57	0,23
5	0,4263	425966,5	74494,06	0,57	1,43	0,25

Robust Regression Results on a validation set

A Robust regression gave better MAE and MAPE scores than linear regression.



Metric	Score
R^2	0.7838
MAE	62977.83
RMSE	141654.72
MAPE	88.52
Average Ratio RMSE	47.42 %
Average Ratio MAE	21.08 %

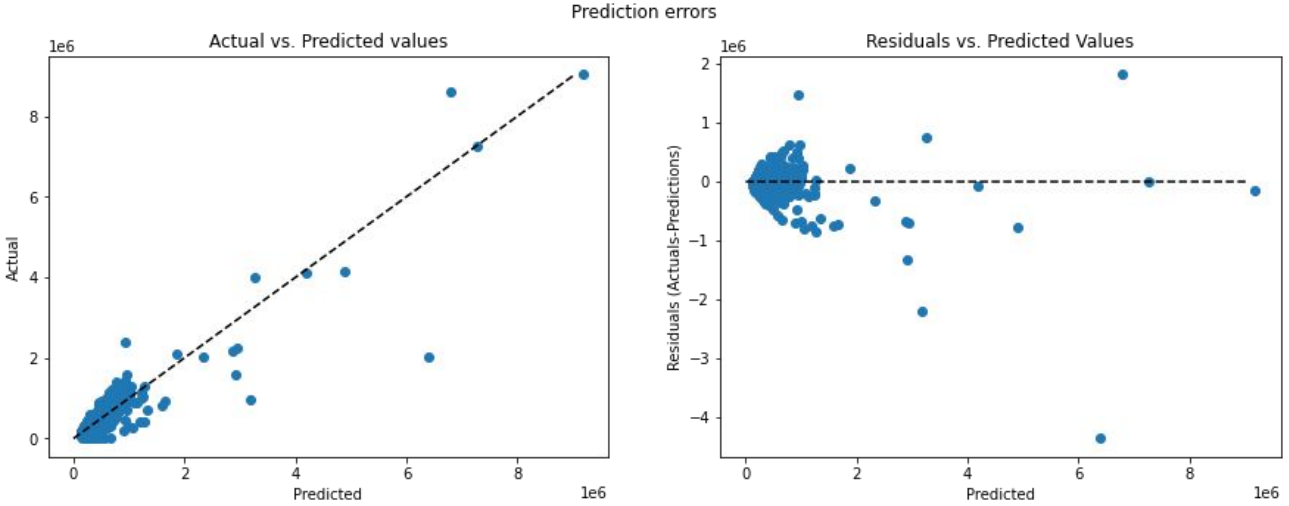
Robust Regression Cross-Validation

Results of Cross-Validation on 5-folds

Fold	R^2	RMSE	MAE	MAPE	Average Ratio RMSE	Average Ratio MAE
1	0,5658	24 9058,4	70 938,52	96,84	0,83	0,24
2	0,7188	13 5996,65	60 897,6	0,39	0,46	0,2
3	0,7106	16 4635,2	65 269,11	69,04	0,55	0,22
4	0,8237	17 5533,08	66 312,7	41,22	0,59	0,22
5	0,4275	42 5490,43	72 969,53	0,51	1,42	0,24

Random Forest Results on a validation set

A Random Forest's results on the validation set.



Metric	Score
R^2	0.8221
MAE	49105.68
RMSE	128461.08
MAPE	91.58
Average Ratio RMSE	43.00 %
Average Ratio MAE	16.43 %

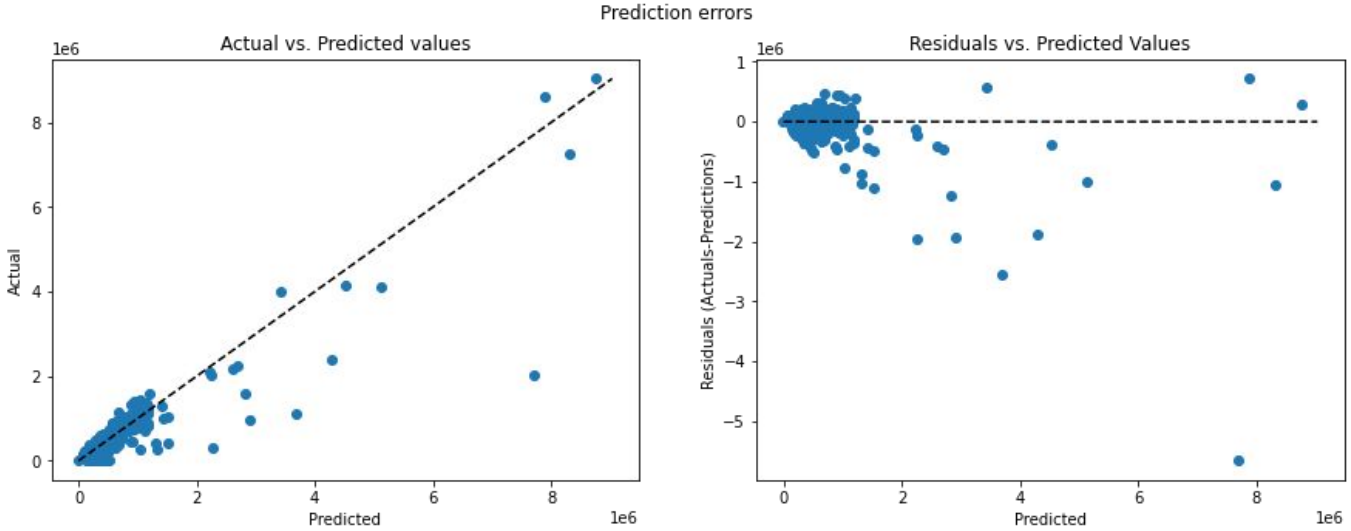
Random Forest Cross-Validation

Results of Cross-Validation on 5-folds

Fold	R^2	RMSE	MAE	MAPE	Average Ratio RMSE	Average Ratio MAE
1	0,6294	23 0119,09	56 218,55	95,15	0,77	0,19
2	0,7902	11 7480,66	47 965,22	0,36	0,39	0,16
3	0,7702	14 6698,54	50 630,49	67,04	0,49	0,17
4	0,8018	18 6118,44	53 462,59	55,64	0,62	0,18
5	0,4388	42 1286,18	58 668,67	0,47	1,41	0,2

XGBoost Results on a validation set

A tuned XGBoost's results on the validation set.



Metric	Score
R^2	0.8102
MAE	33196.17
RMSE	132711.88
MAPE	131.71
Average Ratio RMSE	44.42 %
Average Ratio MAE	11.11 %

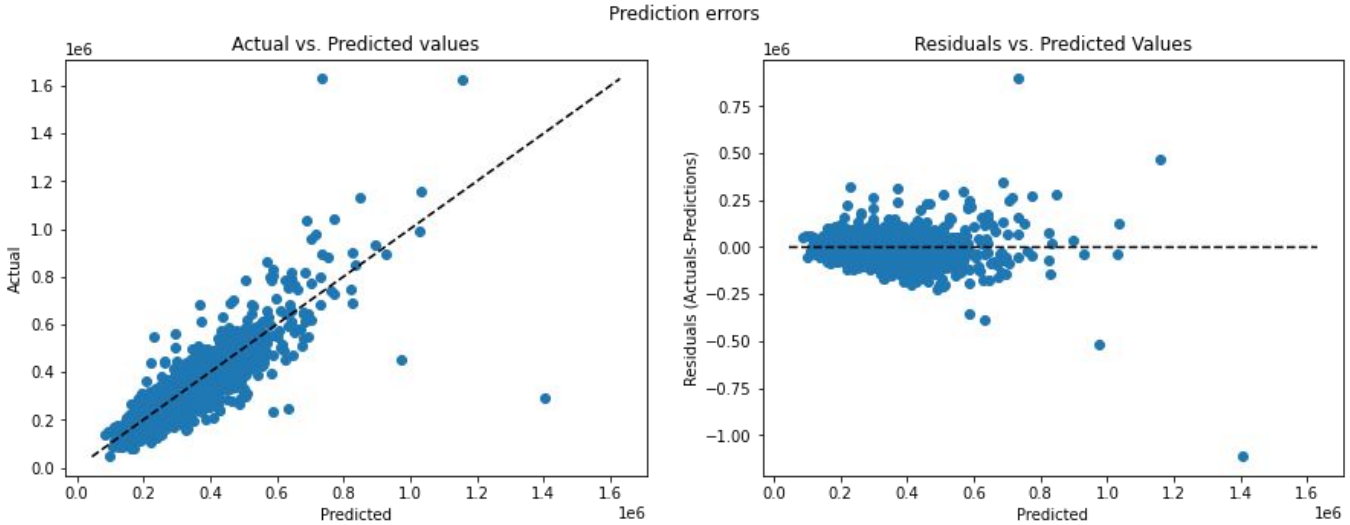
XGBoost Cross-Validation

Results of Cross-Validation on 5-folds

Fold	R^2	RMSE	MAE	MAPE	Average Ratio RMSE	Average Ratio MAE
1	0,5949	24 0565,31	41 697,19	98,94	0,81	0,14
2	0,8068	11 2724,56	33 054,61	0,28	0,38	0,11
3	0,733	15 8124,4	37 797,05	64,69	0,53	0,13
4	0,8095	18 2478,04	40 107,63	54,58	0,61	0,13
5	0,4571	41 4351,25	43 419,45	0,37	1,39	0,15

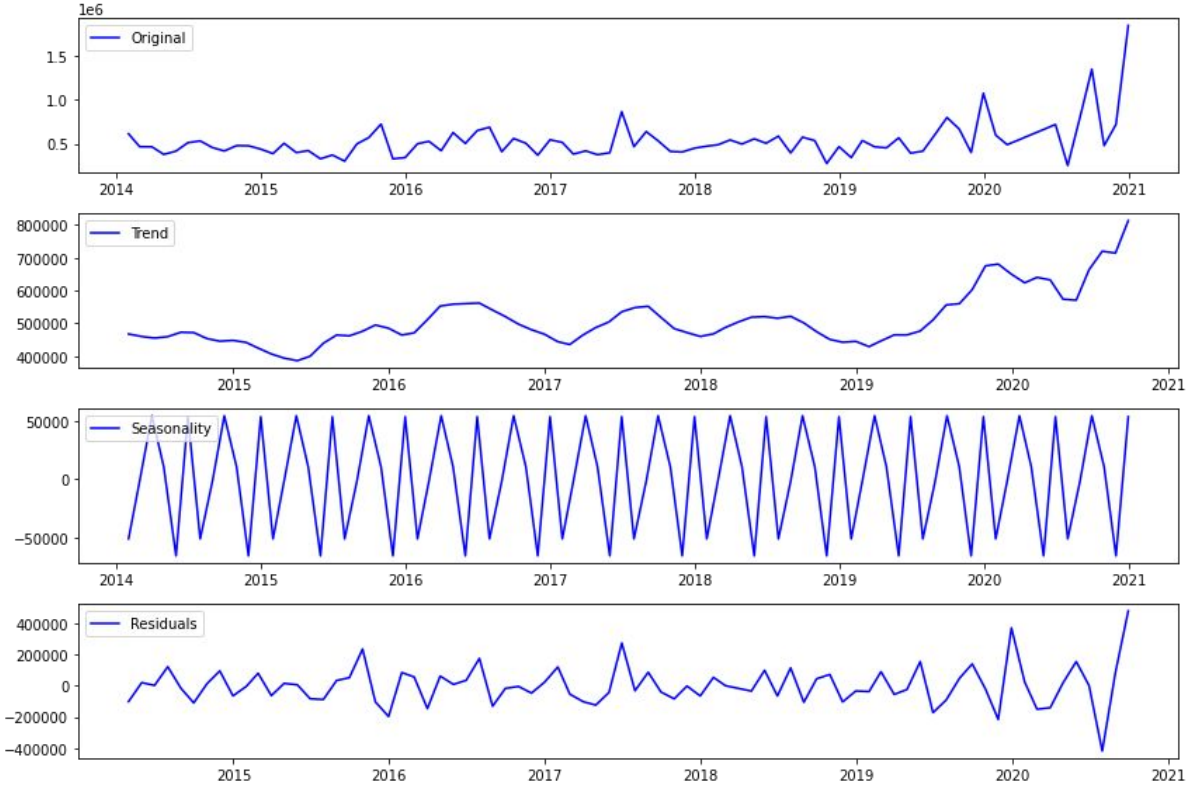
XGBoost Results on TEST set

XGBoost's final results on the test set.

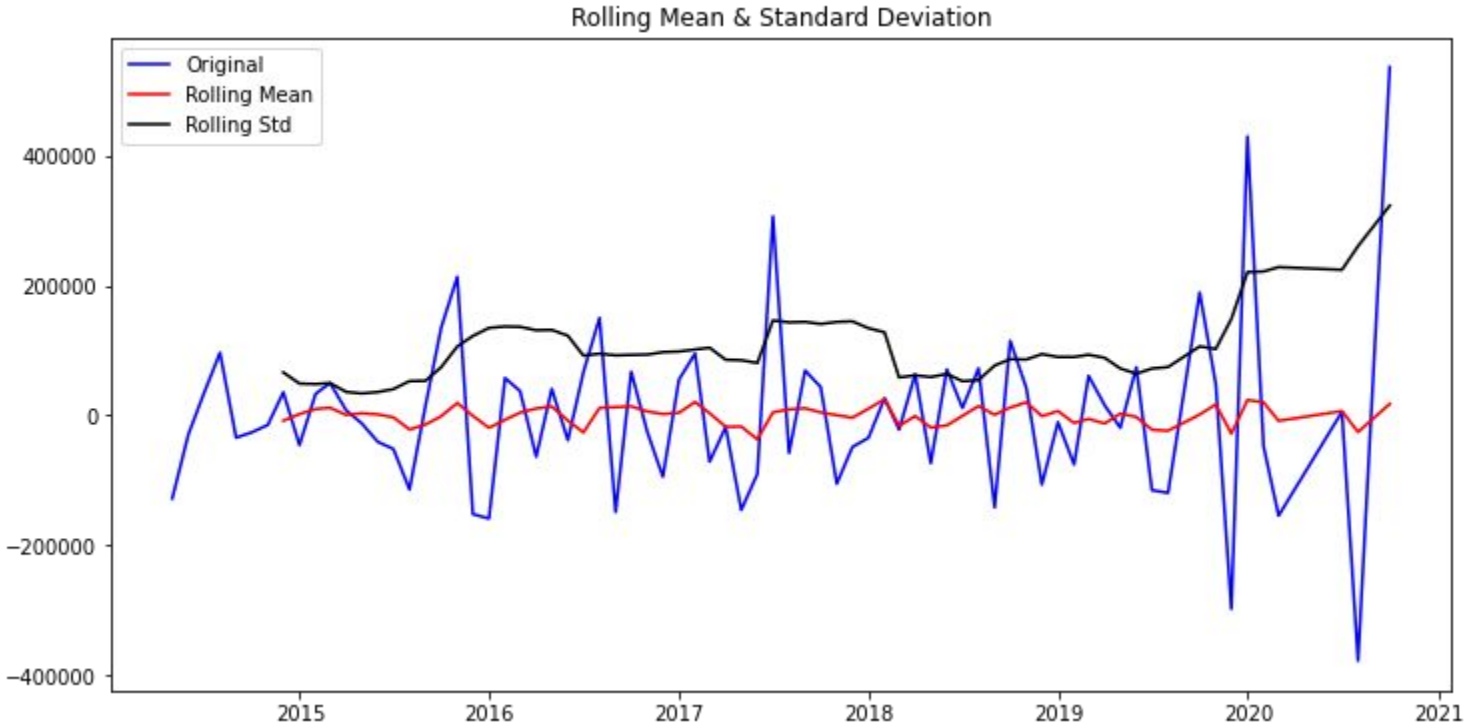


Metric	Score
R^2	0.7740
MAE	35 416.97
RMSE	59480.57
MAPE	0.1357
Average Ratio RMSE	22.41 %
Average Ratio MAE	13.34 %

Time-series Decomposition



Rolling Mean & Standard Deviation



ARIMA models did not seem relevant