REPORT OF ANALYSIS

JULY 22

Authored by: Habiba Mammadli

Analysis of the Sales Prices of New Constructions in Narimanov District

BİNA.AZ site

Dashboard link:
https://public.tableau.com/shared/346TM899F?:display_count=n&:origin=viz_share_link
Github link: https://github.com/habibamammadli/Real_Estate_Analysis_and_Modeling
Agenda:

1) **How can we reduce errors?**
2) **Web-scraping**
3) **Analysis steps**
4) **Modelling**

1. How can we reduce errors?

1) I think that the website description has been written a bit more accurately, and for this reason, during web scraping, we can also scrape the description part and perform text mining with Python code, which can lead to better model accuracy.
2) By scraping the date of the listing, we can obtain information about how long it has been listed on the website.
3) During web scraping, we can also use the number of views of the listing to determine whether it is trending or not, and add it as a feature in the future.
4) We can obtain better results by segmenting and modeling renovated and non-renovated homes using clustering method.

2.Web-scraping

During web scraping, I used the following features for my analysis:

Location - Proximity to the metro.

Price - The full price.

Bill_of_sale - Whether there has been a bill of sale.

Mortgage - Whether it is useful for a mortgage.

Repair - Whether it is repaired or unrepaired.

Area_m2 - The total area.

Number_of_room - The total number of rooms.

Floor - The floor.

I have saved the data I obtained from web scraping into a file named "Task-Dirtydata.csv."

3.Analysis steps

Additionally, I created the "price_per_m2" feature by grouping the data based on location, summing the prices within each group, and then dividing the total price by the total area for the grouped data. This "price_per_m2" feature represents the price per square meter for each location group.

```
Azadlıq Prospekti m.        7
Şah İsmayıl Xətai m.        6
Böyükşor q.                 3
Nizami m.                   3
İçəri Şəhər m.              2
İnşaatçılar m.              2
20 Yanvar m.                1
Sahil m.                    1
Qara Qarayev m.             1
Elmlər Akademiyası m.       1
Ulduz m.                    1
```

I decided to delete some locations that are farther away and less common than our target locations. So, I have deleted the above-mentioned data. After performing statistical analysis, I removed the outliers.

I have examined the correlation between the features and checked whether the distribution of the target variable, which is the "price" column, appears normal in the histogram. I have converted the columns from categorical to numerical to use them in the model later.

In the next step, I have transformed the columns using the Min-Max scaler to achieve a normal distribution.

4. Modelling

1) I have obtained values as follows from the errors of the RandomForestRegressor:

Metrics

MAE 2.478175e+04

MSE 1.288136e+09

RMSE 3.589061e+04

MAE (Mean Absolute Error): This metric measures the average absolute difference between the predicted prices and the actual prices. In this case, an MAE of approximately 24,781.75 AZN suggests that, on average, your model's predictions deviate from the actual prices by this amount.

MSE (Mean Squared Error): MSE measures the average of the squared differences between predicted and actual prices. A high MSE value, in this case, approximately 1.288 billion AZN, indicates larger errors in predictions.

RMSE (Root Mean Squared Error): RMSE is the square root of MSE and provides a measure of the average magnitude of prediction errors. An RMSE of approximately 35,890.61 AZN means that, on average, predictions of model are off by this amount.

2) The results from XGBRegressor model's metrics are as follows:

MAE (Mean Absolute Error): Approximately 23,108.73 AZN

MSE (Mean Squared Error): Approximately 1,198,614,000 AZN

RMSE (Root Mean Squared Error): Approximately 34,621.00 AZN

3) The results from your LGBMRegressor model's metrics are as follows:

MAE (Mean Absolute Error): Approximately 23,220.51 AZN

MSE (Mean Squared Error): Approximately 1,182,998,000 AZN

RMSE (Root Mean Squared Error): Approximately 34,394.74 AZN

Comparing these metrics to the results from the RandomForestRegressor and XGBRegressor models:

The LGBMRegressor model's MAE is close to the XGBRegressor, indicating similar performance in terms of average prediction errors.

The LGBMRegressor model's MSE is also close to the XGBRegressor, suggesting similar performance in terms of squared errors.

The RMSE for the LGBMRegressor is slightly lower than the XGBRegressor, indicating a smaller average magnitude of prediction errors.

Overall, the LGBMRegressor model's performance is similar to that of the XGBRegressor model, with slightly lower RMSE