CRISP-DM Project: Real Estate Price Prediction

Muhammed Esad Mazı

February 2024

1 Business Understanding

1.1 Assessing the Situation and Background Information

Following the global impacts of the COVID-19 pandemic and the Russo-Ukrainian war, the real estate industry has experienced significant growth[1]. The emergence of Gulf countries such as Saudi Arabia, the United Arab Emirates, and Qatar has intensified competition, resulting in a substantial increase in market volume within these nations. Moreover, real estate is a sector fraught with risk, and Saudi Arabia's position in the Middle East adds an additional layer of complexity.

1.2 Business Objectives

In today's dynamic real estate landscape, traditional selling methods are giving way to digital platforms such as aqar.fm. These platforms play a pivotal role for both clients and data analysts, providing a sophisticated avenue for property transactions.
Our client seeks guidance in predicting housing prices in Kingdom of Saudi Arabia using housing characteristics. However, making this decision is far from straightforward. The business goals have been collaboratively established with various departments and our client. Our main goal is:

- ”Given the qualities of a home, what is the predicted price? (Only sells)”

1.3 Data Mining Goals

The primary objective of this project is to develop a predictive model that empowers our client to determine accurate housing prices through comprehensive data analysis.

- Build a model with reasonably good RMSE value(less than 50K SAR).

2 Data Understanding

2.1 Data Collection

The database we are going to use is the Saudi Arabia Real Estate dataset, scraped from real-estate website aqar.fm,one of the major online real estate websites.

2.2 Data Description

Our Data was in .db format, we transformed it to .csv format. This database consist of 48 columns and more than 663K rows. Examples of features can be given as: id,price, category, kitchen furnished etc. Some features and their descriptions are given in the table 1 below. To show it better, there are some examples from data:

Feature	**Description**
user_id	ID of user
price	Price in Saudi Riyal(SAR)
refresh	The time of last refresh in UNIX time
wc	# of toilets
city	City of the listing

Table 1: Description of some Features

Feature	**Example 1**	**Example 2**	**Example 3**
user_id	10065	11005	8017
price	1958400.0	3000000.0	8057700.0
refresh	1672495809	1672040788	1671819056
wc	0.0	5.0	NaN
city	Riyadh	Madinah	Jeddah

Table 2: Examples from Data

Our table has a lot of missing or irrelevant values, outliers and useless columns. In Figure 1, we can clearly see number of missing values in each column. To increase readability, we will filter out the columns which have less than 330K(50%) NaN/empty values. in Data Preparation part.

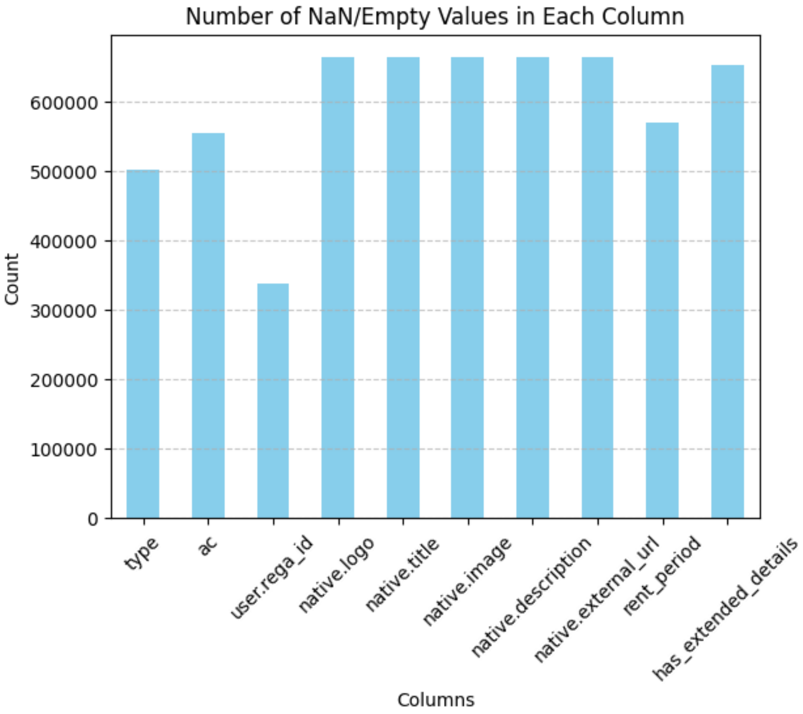

Figure 1: Number of NaN/Empty Values in Each Column(Filtered)

Using our category feature, we can split the dataset into rentals and sells. After splitting, we can see that we have mostly sell listings(527405 entries) compared to rental listings(136541 entries).

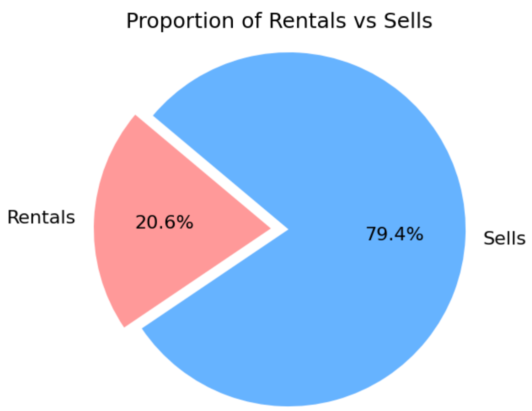

Figure 2: Proportion of Rentals vs Sells

To visualize data further, we use histograms. We included most meaningful histograms to increase readability.

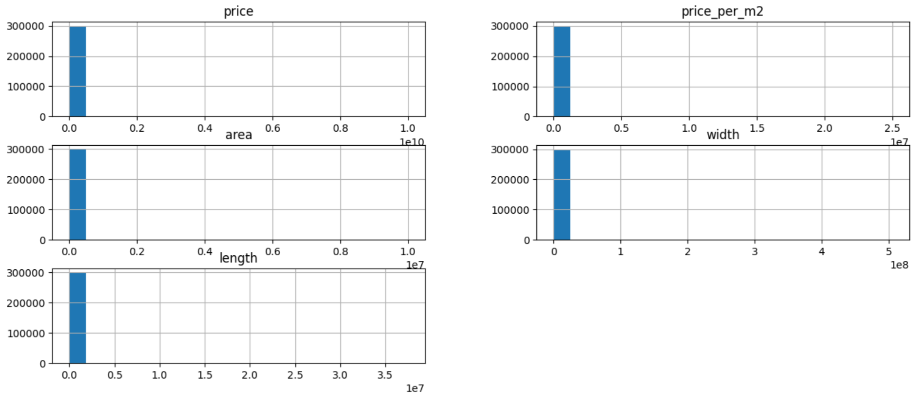

Figure 3: Histograms of some features.

As we can clearly see, our data is very left skewed due to outliers. we will to make it closer to Gaussian distribution by eliminating outliers.

2.3 Data Quality

Our data is raw and needs work before usage. We need to select, clean, construct and format data in the next part according to CRISP-DM methodology.

3 Data Preparation

3.1 Data Selection

We need to select what we need to accomplish our goal. That's why we need to include some columns and exclude others.

We Exclude:

- id,refresh,last_update,district_id,city_id,create_time,createdAt,updatedAt, user_id, uri, title, content, imgs, path, user.name, user.phone, user.img **because they are not related to our goals.**
- type, ketchen, ac, rent_period, has_extended_details, daily_rentable **because they are mostly consist of NaN/empty cells.**

We include all other 19 features.

3.2 Data Cleaning

Do clear data, we delete some rows with NaN features. Here is the list and number of features we delete:

Feature	**# NaN Rows**
area	577
street_width	18420
user.review	113586
user.iam_verified	3410
advertiser_type	18152
street_direction	71135
width	41433
length	218

Table 3: Number of Deleted Rows

Moreover, we delete all rows that has "0" for user.iam_verified because we want to work on only verified users for more accuracy. We can d After this cleaning process, we got 509801 rows instead of 393545.

We also want to work only on sells listings, not rentals. For this, we use category feature to clean the data further. At this point, we have 275373 rows.

Since our data has a lot of Missing Not at Random (MNAR) values, it is a challenge to assign correct value for our NaN values in order to not lose information and create bias in our model.

- We tally the number of missing attributes ('age', 'beds', 'wc', 'living') for each row in the dataset. This information is stored in a new column called 'missing_count'
- Next, we filter the dataset to retain rows where either all 4 attributes are missing or all 4 are present.
- After filtering, we discard the 'missing_count' column as it's no longer needed for further analysis.
- We will create a new feature: if these 4 features are NaN,then this new feature will be 1. otherwise, it will be 0. We will call this feature land type.

3.3 Data Construction

To make a more detailed and better model we need to construct new features from what we have. Here are the new features and their descriptions:

New Feature	**Description**
total_rooms	# of beds,wc,livings in Total
price_per_m2	Price per m^2
dist_from_center	Distance from City Center in Kilometers

Table 4: Newly Constructed Features

We created "dist_from_center" feature using OpenStreetMap Nominatim API[1]. We calculated location using "location.lat" and "location.lng" features and locations from the API.

We removed outliers with 95% to make our data more normally distributed. That way our model will respond much better.

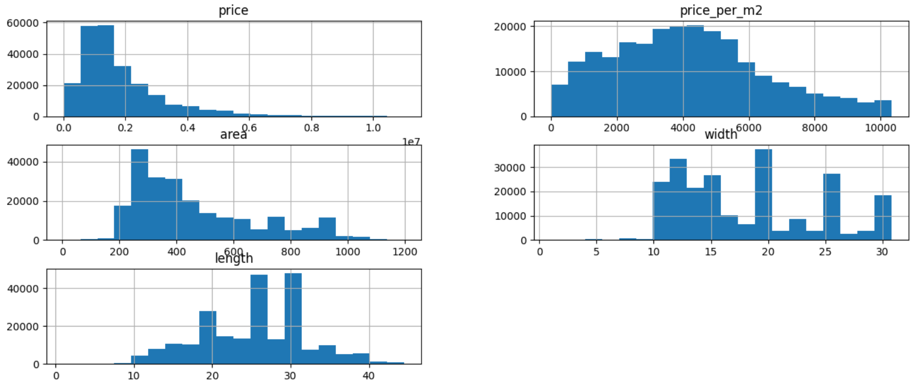

Figure 4: Histograms of Some Features without Outliers

Moreover, we one-hot encoded the category feature to ensures that our model does not assume that higher numbers are more important in our category feature. After this, we delete the category feature. We also one-hot encode the advertiser_type feature for same reasons.

For our city and district features, we used frequency encoding since they have a lot unique values.

Furthermore, our endeavor into time series analysis led us to a formal conclusion: regrettably, our dataset demonstrated a lack of suitability for deriving meaningful insights through this methodological approach.

3.4 Data Integration

At this point, we have not integrated any data. Our project is open to further changes.

3.5 Data Formatting

We format data by renaming these features for more readability and functionality:

- location.lat → location_lat
- location.lng → location_lng
- user.review → user_review
- user.iam_verified → user_iam_verified

4 Modeling

4.1 Select Modeling Technique

We have not choose a model. To pick most useful model, we need to test different models with a sample. We will consider "Linear Regression", "Decision Tree", "Random Forest", "Gradient Boosting" and "Support Vector Machine" models from scikit-learn library[2].

After we test these models with a sample size of 20000 and cross-validation, we found out that best model is "Random Forest".

4.2 Generate Test Design

It is very important to note that before building the model we need to create test design. In prediction tasks such as ours, it is common to use errors as a metric. We will also use errors to assess scores for our model. We will not use validation set and we will split our data into 80/20 split as usual.

4.3 Build Model

We ran the modeling tool RandomForestRegressor()[2]. Our model gave these results:

```
Test RMSE: 42802.22529023477
```

In our model, we achieved excellent results with a Root Mean Squared Error (RMSE) of 42802.23. This metric indicates that, on average, our model's predictions deviate by approximately 42802.23 SAR from the actual housing prices. Given the scale of housing prices, this level of accuracy is highly desirable, suggesting that our model makes predictions with relatively low error.

Overall, this outstanding RMSE value underscores the effectiveness of our regression model in predicting housing prices. It provides strong evidence of the model's reliability and accuracy, making it a valuable tool for housing price prediction tasks.

4.4 Assess Model

With further discuss with our business team, we concluded that our model meets our business and data mining goals. We stop investing more on our model further because it already gave enough performance.

5 Evaluation

5.1 Evaluation of Results

In this section, we evaluate the results of our data mining project, which focused on predicting housing prices in Saudi Arabia. The evaluation is framed around how well the predictive model meets the set business objectives and whether any deficiencies exist for business reasons.

5.1.1 Assessment Against Business Objectives

Our project set out with clear business objectives to predict housing prices using detailed housing characteristics. The final model chosen, a RandomForestRegressor, demonstrated high accuracy with a Root Mean Squared Error (RMSE) of 42,802.23 SAR. This indicates a strong performance as the deviations from actual prices are relatively low, suggesting that the model is reliable and provides valuable predictions for our client.

5.1.2 Novelty and Utility of Findings

Apart from the models, our findings include identifying the key features that impact housing prices, such as location, area, and amenities provided. These findings are crucial as they not only help in enhancing model accuracy but also assist real estate stakeholders to understand market dynamics better.

5.1.3 Comparison with Initial Evaluation Criteria

Initially, we aimed to achieve a model accuracy 50K SAD RMSE. Our chosen model's performance, indicated by the RMSE, suggests that we have exceeded these expectations. The effectiveness of the model in real-world applications would further validate these metrics.

5.1.4 Additional Insights

The data mining exercise also uncovered significant data quality issues and outliers, which were addressed during the data preparation phase. These insights are valuable for the client as they highlight the importance of data management and could lead to better data governance practices.

5.1.5 Approval and Ranking of Models

Based on the evaluation against business success criteria, the RandomForestRegressor model was approved for deployment. It meets the required criteria effectively, indicating that further investment in model refinement may not be necessary at this stage.

5.1.6 Conclusion

The evaluation of the data mining project confirms that the results align well with the business objectives. The model not only predicts housing prices with high accuracy but also offers insights that are beneficial beyond the immediate scope of price prediction. This dual output of models and findings adheres to our defined equation: RESULTS = MODELS + FINDINGS, successfully capturing the holistic output of our data mining efforts.

5.2 Review Process

The review process serves as a quality assurance step, ensuring the model's efficacy and alignment with business needs. At this juncture, the model appears to satisfy the set business objectives, prompting a thorough examination of the entire data mining engagement to identify any overlooked or potential improvements.

5.2.1 Analysis of the DM Process

In our Data Mining processes, there where a lot of points which were potential pitfalls. Since our data was not perfect, we spend quality time and effort on Data Understanding and Data Preparation part.

A possible failure point was setting Business Objective for our clients too high or unachievable. Since our team is well experienced, we knew that anything can go unexpected. Hence, we set our business objective relatively flexible in the beginning. After understanding what we have as data and what can we do with it, we returned our business objective part.

In our project, there were many misleading steps we observed. Thankfully we noticed them before and acted upon them. For example, when we want to use time series in our model, we noticed that our features related to time/date features were actually meaningless corrupted data. Also, overfitting of the model was a thing in the beginning. Our team mitigate overfitting by returning to data preparation part again and again.

5.2.2 Conclusion

This comprehensive review is critical for ensuring the project's success, providing a foundation for future enhancements and confirming the model's readiness for deployment or further refinement.

5.3 Determine Next Steps

Following the comprehensive evaluation and the process review, our project team is at a crucial decision-making point. We must determine the most effective course of action to either move forward with the deployment of our results, initiate further iterations for process improvement, or embark on new data mining projects that build on our current insights.

The analysis of the potential for deploying each result has shown immediate business value and could provide quick returns on investment. However, there remains a risk if these models underperform with new, unseen data. Estimating the potential for improvement in the current process suggests significant benefits such as enhanced accuracy and robustness of the models, yet this requires additional time and resources which could delay the realization of benefits.

Moreover, a check on remaining resources indicates that while further refinement is possible, it could potentially divert attention and resources from other essential business needs. Recommending alternative continuations could open new avenues for innovation and exploration, aligning with the dynamic nature of our industry, but this would also demand new strategic focuses and investments.

After careful deliberation, the team has decided to prioritize the estimation of potential for improvement of the current process. This decision is driven by the goal to maximize the effectiveness and robustness of our models before full-scale deployment, ensuring that they are as reliable and accurate as possible. This approach not only addresses immediate project needs but also sets a precedent for rigorous quality assurance in future data mining endeavors.

This strategic decision is documented with a clear rationale, emphasizing a balanced approach between achieving quick wins and ensuring long-term project success. By enhancing the current models, we position the project to deliver sustained value to the business, supporting a foundation for future initiatives and ongoing improvements.

References

[1] OpenStreetMap contributors. OpenStreetMap Nominatim API. Software, 2024.

[2] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.