CRISP-DM analysis of Saudi Arabia Real Estate dataset scraped from aqar.fm

Polina Zelenskaya
Innopolis University
Innopolis, Russia
p.zelenskaya@innopolis.university

***Abstract*—Aqar is popular platform to rent and sell real estate in Saudi Arabia. In this paper we will apply CRISP-DM analysis on their real estate dataset in order to identify possible beneficial use cases of data.**

***Index Terms*—CRISP-DM, Real Estate, Analytic, Machine Learning.**

I. Introduction

CCRoss-Industry Standard Process for Data Mining (CRISP-DM) is a standardized method to analyze, describe and research data for industrial purposes. This technique aims to make discovered results reliable and repeatable, as well as comprehensible by people with little-to-no Data Mining (DM) experience. This approach consists of 6 major steps and, if done correctly, leads to an in-depth understanding of data and use cases it could either benefit or harm.

Saudi Arabia Real Estate dataset [1] contains information about real estate in Saudi Arabia scraped from Aqar [2] online platform. Data consist of text descriptions, photos, locations and numeric values representing prices, phone numbers e.t.c. In total it contains 45 different features with 663946 samples.

II. Business Understanding

A. Business Objectives

To start the analysis, it is important to understand business situation itself (Due to language barrier and translation difficulties, this part should be considered carefully, as all information is taken from one official website [2] and post-proceeded by google's Arabic-English translator). Aqar is the largest real estate platform in Saudi Arabia. It attracts 50 million viewers and contain information of about 1.5 million properties. Company mostly operates in digital space and makes money from selling leads to real estate agents, selling add spaces and recommendation priority, as well as selling subscription models for customers. According to [3] Company's net worth is estimated around 7.9 million dollars as of 2019.

Considering everything have said, we can now form business objective. From business perspective, clients primarily want an easy to use platform to find commodities with reasonable price, truthfully description and safety guarantees. However, Aqar also provides interface for real estate agents, so from their point of view, it is crucial that their advertisement will be shown to a target audience and result in a high engagement rate. And additional objectives are platform stability, high priority in search engines, client market expansion and client return rate.

It is also important to state business success criteria. As we are not related to company directly and have no insights about their company culture and goals, we can only speculate. Success could be measured as number of new users signing in month, number of active users on the platform, total value of transactions between real-estate agents and clients, retention rate or just revenue generated.

B. Situation Assessment

In order to conduct this analysis we need to consider available resources. First of all we have Saudi Arabia Real Estate dataset [1] with over 300 thousand objects each having 45 features, official Aqar website [2] and Arabic-English google translator to break language barrier. Speaking about experts, we are free to discuss this work with our Teacher Assistant Alaa Aldin Hajjar and Professor Armen Beklaryan. In discussion of computational resources, we have Nvidia 2080 SUPER mobile GPU and Intel core i9 processor, this should be enough to training light models, and scrape information in reasonable amount of time. Considering our team knowledge, we will use Python 3 programming language and frameworks it offers (Scipy, Pandas, Numpy, Matplotlib, Pytorch, Requests, BeautifulSoup4).

It is also important to discuss requirements, assumptions and constrains. There are several report requirements: work deadline is 13th of May 2024 (work started at 27th of January 2024), and analysis in a result should provide beneficial solution to company, with highlighted data patterns ans outcomes. Speaking about data assumptions, we suppose that there are patterns in datataset and those patters correlate with business success. In addition, we can suppose that data is fair and unbiased, as well as have no crucial mistakes (such as invalid locations, unreasonable prices e.t.c.). And discussing constraints, we should note that language barrier is the main one, as no one in our team know Arabic. In addition, we have no connection with the Aqar company, so all company details is from public resources found on the internet. Dataset quality is data constrain, as we use already scraped data and we do not have resources to collect additional information, this is especially important as dataset covers only dates from 01.01.2023 to 04.08.2023 (so it is already outdated for half a year).

Risks are not business related, as this is educational project and only our success on the course is influenced by it. However if it will be deployed, ethical considerations should be made, model should be tested on different kind of bias it could face. In addition, if model won't be provided new data it may get outdated and remove all benefits it was intended to add. For safety reasons, explainable and simple approach should be developed, that will be enabled instead of model in case of emergency.

Now we can discuss cost and benefits. From this dataset we see potential from price prediction model, it could provide new functionality to platform, be used in a subscriptions for more in-depth analysis of a price customers see, and show real estate agents what price is reasonable for customers to buy or rent property. Also if model will support opportunity to an importance feature matrix, we can show clients and agent how features influence the price. Now we can dive into details. According to [3], company have average of 175 thousand unique visitors each day, and generates about $4300\$$ daily from ads. If we want to embeed price prediction model, we will need to add dedicated server with modern GPU to support such amount of clients. We can suppose that most of the server requests come in a daytime for Saudi Arabia (their target clients are people from Saudi Arabia) and more computational power required during day. For that reason, to suppose average number of requests we will divide total number of unique users by number of working hours (8) and result in approximately $6 \approx \frac{175673}{60 \cdot 60 \cdot 8}$ new unique users each second. As we need to estimate upper bound, we can suppose that each user will spend on average 2 hour on the website, resulting in $44000 \approx \frac{175673}{60 \cdot 60 \cdot 8} \cdot (2 \cdot 60 \cdot 60)$ users at the same time online, each sending requests to our new dedicated server. For such load we will require a high performance computing server, which we can rent with Tesla T4 cluster from Yandex Clound [4] for around $950\$$ a month. With this in mind, we now need to determine how model could benefit the company. If it will be introduced as a subscription to provide enhanced information about price and feature correlations, to fulfill the expenses we can determine number of subscribed clients required if the month subscription price will be p - $N = \frac{950}{p}$. We can see that even with $1\$$ payment we only need 950 subscribed clients, which is reasonable amount considering daily count of visitors is around $175k$. With caching system we can cut down costs a lot more, however we will only consider upper bound in this case. Speaking about benefits from this model, we can't say for sure number of recurring users, or number of new users visiting the platform daily, however we have information that daily platform faces $175k$ visitors, and according to [5] industry average of all users willing to buy subscriptions is 0.62%. Knowing all that we can make rough estimate that approximately $1085 \approx 175673 \cdot 0.0062$ users would get a subscription. And with price as low as 1 per month, this model would support itself in make small revenue in the long run, and with $p = 5\$$ it could boost daily income from estimated [3] $4321\$$ to $4470\$ \approx 4321\$ - \frac{950\$}{30} + \frac{1085 \cdot 5\$}{30}$.

In addition to production costs, we also need to consider cost to develop this project. From data analysis point of view, this project could be done by one Data Science Junior developer. The time constraint is difficult to estimate as it could vary from individual experience, however considering our project plan, we suppose that this work could be done and fine-tuned in 3 month. With this in mind we can estimate that developing expenses. According to [6], average Junior developer salary in Saudi Arabia is about $50000\$$ per year, meaning 3 month of development will cost approximately $12500\$ \approx 3 \cdot \frac{50000\$}{12}$. With $p = 5\$$, company will profit from project in 84 days $84 \geq 3 \cdot \frac{50000\$}{12} \cdot \frac{1}{4470\$ - 4321\$}$, and with $p = 1\$$ in around 2778 days $2778 \geq 3 \cdot \frac{50000\$}{12} \cdot \frac{1}{4325.5\$ - 4321\$}$.

C. Data Mining Goals

Now when we estimated cost benefits, we can describe our data mining goals. As we already discussed in *Situation assessment*, main goal is to develop a model that will predict price and show how each real estate features correlate to it. The most important factor is model's interoperability, so deep learning technique should be considered carefully.

Considering model's evaluation, in this problem MSE metric should be preffered to MAE as one mistake in $1000\$$ is more crucial than 10 mistakes by $100\$$. Model always predicting in the correct hundreds would be considered good, but even more precise predictions would benefit even more. Speaking about subjective evaluation, model should be unbiased and provide results that would correlate with human's intuition.

D. Project Plan

In order to develop this project we need to consider each CRISP-DM step. In total there are 6 steps (Business Understanding, Data understanding, Data preparation, Modeling, Evaluating, Deploying), where on each step we have 2 weeks to develop and report. As 01.02.2024 is deadline for first step, we can calculate each new deadline accordingly. For the last two weeks deadline will be report finalization and outcomes of the work.

Tools we will be using is Python3 with machine learning stack (Pytorch, Matplotlib, Gradio, Sklearn, Pandas, Scipy, Numpy). If required, we will consider some additional libraries and tools.

III. Data Understanding

A. Collect Initial Data

First of all, let us discuss the collection step. As we decided to use [1], which is Saudi Arabia Real Estate dataset that contains information about buy and rent price of properties throughout the whole county. This data is already scrapped and all information in the dataset is taken from [2]. For now, we consider this dataset sufficient to make prediction model, as all features are taken directly from Aqar. This means that new data appearing after deployment would be treated identically with no additional scraping required, which simplifies both deployment and maintainment of the model.

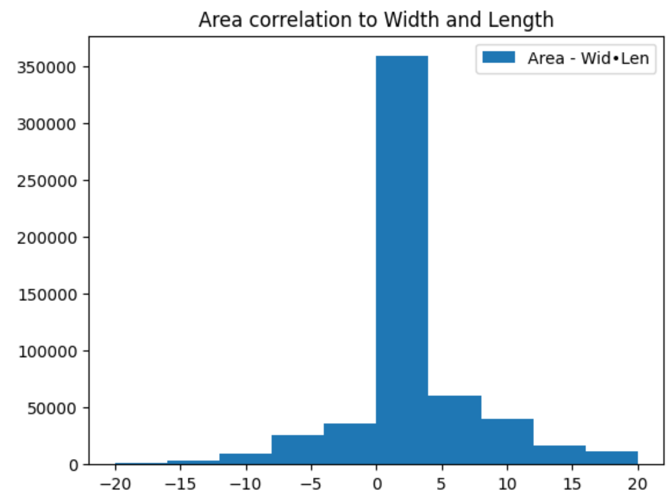

Fig. 1: Deviation of ‘Area‘ and it’s estimation via ‘Length‘ and ‘Width‘ on -20 to 20 range with 10 bins.

All statistics from this section is available publicly on our GitHub repository [1] in ’Data exploration’ notebook.

B. Data Description

Considering this data superficially, we can say that [1] contains of 45 different features with 663946 samples. Features contains information about house location, price, inside structure (number of rooms and their types), create time, user reviews, owner information, estate category (rental or sell, apartment or villa or room) e.t.c. In addition, it contains images, which could also be used. All this information is taken from [2], which is convenient as it would be simple to increase dataset if required. Worth mentioning data modernity. This dataset consist information for time-period from $01.01.2023$ to $04.08.2023$, and as for year 2024 it is already outdated by half a year.

C. Data Exploration

First of all, let us discuss target - price. It has mean of 3986665 and std of 62561260. Most interesting part is that prices mostly located at low values, and then gradually decrease as cost rises. In addition to that, there is spike for less than 5000 values, probably indicating rent costs.

Speaking about some feature correlations, we decided to investigate how ‘length‘ and ‘width‘ correlates with ‘area‘. For that reason, we applied basic idea that ‘area‘ is approximately ‘length‘ times ‘width. To proof this, we subtracted from ‘length‘ product of ‘length‘ and ‘width‘, and surprisingly enough, this estimate shows strong correlation between these three columns. With error window ranging from -100 to 100 we cover 97% of data, with range from -15 to 15 we cover 90%, and with range as small as 0 over 50% of samples. Whole data could be seen in Fig. 1. For this strong correlation, we will consider deleting these columns to remove multicollinearity.

Discussing real estate owners, mean rating is 4.41. Interestingly, 25 percentile is equals to 4.19, meaning most of owners have rating of at least 4. Around 99% of users verified, and about half have ‘rega_id‘ filled. Around 34.9% are classified as ‘normal_marketer‘, 30.9% ‘exclusive_marketer‘, 19.6% ‘agents‘ and 10.8% ‘owners‘.

Most of the real estate is located in the capital of Saudi Arabia Riyadh (53.2%), followed by Jeddah (15.3%), Dammam (5.5%) and Al Khobar (4.6%). Popular distincts are An Narjis (5.1%) and Al Aarid (4.1%). Around half (48.4%) of the real estate was posted in the January of 2023. Average house contains 3 wc, 1 living room and 4 beds. About 58.5% of real estate is not furnished, and around 53.9% having kitchen. Most of the estate promoted being new with 38.5% was built less than a year ago and mean age value of 3.78 years.

D. Data Quality Verification

First of all, before going in-depth, we can look into how pure our data are. By looking at number of missed values, we can see that all data named ‘native.*‘ (5 columns) are almost fully absent (with only 2 samples containing any information about them). For that reason, we consider that these columns have no important information, so we can remove them entirely.

In addition, column ‘has_extended_details‘ has only 10422 (1.5%) non-Nan values. However, this feature is binary and all filled values equals to 1, meaning we can fill this column by 0 to restore all data.

Worth mentioning, column ‘rent_period‘ also seems to contain a lot Nan’s, to be precise 571232 (86%) samples. Firstly this information seems to be logical as sales advertisements won’t have information about rent period, however if we will consider ‘rent_period‘ column for any property for sale (e.x. ‘class‘ 2), we see that some of them will have this information filled. Meaning not only that rent period do not indicate rental only adverts, but also that some further filtering of the data is a must. Furthermore, there exists ‘daily_rentable‘ column with 315493 (47.5%) missing values, which is also present in properties for sale and have all kinds of ‘rent_period‘ values.

Speaking about interior features, almost third of all values missed in ‘furnished‘, ‘livings‘, ‘ketchen‘, ‘wc‘ and ‘beds‘, and 555898 (83%) values missed in ‘ac‘ column. It is hard to say for sure whether this information is not filled on purpose or by mistake, so we will not impute these columns.

Discussing exterior features situation is better. The most unfilled column is ‘age‘ with about third of values being Nan, and for the ‘street_width‘, ‘street_direction‘, ‘width‘ and ‘length‘ on average 8% values missed.

Considering home holder themselves, 118066 (17.7%) owners do not have reviews, 222565 (33.5%) do not have profile image, and 338325 (50.9%) did not provide importation about their Real Estate General Authority id.

And lastly, features ‘type‘ and ‘advertiser_type‘ have 502773 (75%) and 23932 (3%) respectively. The main problem with these features is that they are hard to interpret, as no description in dataset itself [1] or Aqar platform [2] could be

[1] https://github.com/cutefluffyfox/saudi-arabia-real-estate

found. For that reason, we consider to remove such features in the future.

References

[1] Mohd PH. 2023. Saudi Arabia Real Estate dataset. https://www.kaggle.com/datasets/mohdph/saudi-arabia-real-estate-dataset/data .

[2] Aqar - Platform for real-estate marketing. 2014. https://sa.aqar.fm/ .

[3] SiteIndices - Ranking platform that collects information about business websites and estimates businesses value and income. 2019. https://aqar.fm.siteindices.com/ .

[4] Yandex Cloud - Cloud platform providing scalable infrastructure and storage, for digital services and applications. 2024. https://cloud.yandex.com/en-ru/prices .

[5] Service Form - advertisement company that helps growing business through internet. 2023. https://www.serviceform.com/blog/ultimate-guide-to-real-estate-website-conversion-marketing .

[6] SalaryExpert - publication platform of Economic Research Institute studies. 1989. https://www.salaryexpert.com/salary/job/data-scientist/saudi-arabia .