Real Estate Market Analysis Tool

Supervisor: Fairouz Medjahed Ph.D.

Team Members

Joe Farah

Samia Mahdaoui

Contents

Overview

The proposed capstone project, titled "Real Estate Market Analysis Tool," aims to create an innovative digital tool (a website) that leverages the precision and speed of AI algorithms to conduct in-depth analyses of the highly dynamic real estate market. This task is particularly timely, given the current rapid growth in the real estate sector, presenting promising opportunities for both investors and brokers.

In today's fast-evolving real estate market, data-driven decision-making has gained utmost importance. Investors and industry professionals increasingly rely on data analytics to identify potential opportunities and make well-informed choices. For instance, an article titled "How to Predict Real Estate Prices" highlights that real estate agents consider various market indicators to predict property prices. One crucial indicator they take into account is "Showing Requests," which signifies the level of interest and demand from potential buyers. Essentially, real estate agents must examine a large volume of property listings and show requests to make accurate predictions about market trends.

However, this process is not only time-consuming but also inefficient when done manually. This inefficiency highlights the significant potential for improvement through the application of AI and ML technologies. By automating the analysis of real estate market data, our project aims to streamline and enhance the decision-making process for real estate professionals, making it faster, more accurate, and ultimately more effective.

Project description

Goals

In this project, our aim is to provide an AI based data driven tool (or website) for real estate professionals to make accurate property price predictions in the real estate market.

Motivations

One of the primary motivations driving this project is the pressing need for efficient and data-driven decision-making in the rapidly evolving real estate market. As stated in the article "Machine learning and artificial intelligence in a real estate marketing: a systematic review" (Bykovskii & Dutot, 2022):

"Automation of these processes, supported by collecting and analyzing big data, ensures the marketing practice to be improved, but the prospective opportunities of using Artificial Intelligence (AI) have not been revealed especially in complex personalization tasks."

The challenge posed by the large volume of data points to an opportunity for innovation and improvement. This drives our motivation to utilize the immense capabilities of AI, with the aim of providing real estate professionals with enhanced access to valuable insights.

On the other hand, real estate agents have to go through a large number of listings in order to make property price decisions. Time is the greatest commodity, and this is also true in the real estate realm. As stated in the article "How to predict real

estate prices", real estate agents have to sift through thousands if not millions of property listings. This has proven to be time consuming and inefficient, which is why we believe that AI can benefit these fields. AI models have the ability to analyze and go over a large number of datasets, which will provide real estate professionals with up-to-the-minute insights to streamline data collection and analysis, ensuring they can seize opportunities promptly and effectively.

Machine learning models and data sets

In this project, we will implement and train a neural network-based regression model with the objective of predicting property prices. This model will utilize a feature set that combines both structured and unstructured data, ensuring a holistic approach to price predictions. Achieving optimal performance is crucial, and we will make sure to fine-tune the model, taking into account the computational demands and intricacies of the data. This fine-tuning process may involve iterative training and adjustments.

Furthermore, we will use a pre-existing NLP model, potentially employing a transformer architecture like BERT or GPT-3. This NLP model will transform cleaned review texts into semantically-rich embeddings (using tokenization), effectively converting each review into a high-dimensional vector. This step plays a crucial role in representing the unstructured textual data associated with each property listing.

To support these efforts, we will scrape and acquire a dataset from https://emirates.estate, encompassing key attributes such as property size, bedroom count, house size, bathroom count, property type, and location quality. Within this data collection process, we will also perform essential preprocessing tasks on the reviews, including punctuation removal, conversion to lowercase, stop word removal, and

lemmatization. These comprehensive steps will contribute to the development of a robust and data-driven approach to real estate price prediction.

However, there's a possibility that our neural network-based regression model may achieve a low accuracy score in predicting property prices. Such an outcome could limit the usefulness of the application for real estate professionals, potentially affecting its adoption and overall success in the market.

Risks & Challenges

During the brainstorming around this project, we were able to identify the following potential risks and constraints:

1. **Inadequate Data Availability:** Another concern is the availability of sufficient and relevant data. We may encounter challenges in acquiring an extensive and diverse dataset from https://emirates.estate, which could hinder our ability to create a robust model and generate reliable and accurate predictions.
2. **Website Scraping Policies:** Some websites have strict scraping policies, meaning it is illegal to scrape their information and this may arise as a potential challenge, which may hinder our ability to acquire the necessary data for training, cross-validation and testing purposes.
3. **Scarcity of Reviews on Listing Websites:** The scarcity of reviews on real estate listing websites poses a potential challenge. If there is insufficient amount of review data for each property listing, it may limit our capacity to harness the power of NLP models for text-based analysis and could lead to less informative embeddings. In the case where there are no meaningful reviews, the option will be to completely disregard this feature in the project.

4. **Low Computational Power:** The computational demands of training and fine-tuning complex neural networks, especially when dealing with large datasets, could stress our available computational resources. Inadequate computational power might result in prolonged training times or limit the complexity of our models.
5. **Scope Definition Errors:** An error in defining the project scope is also a risk to consider. Misaligned project objectives or unclear requirements may lead to deviations from the intended goals, potentially impacting the project's overall success and efficiency.
6. **Failure to Reach the Goal:** Ultimately, the most significant risk is the possibility of failing to reach our intended goal of creating a valuable real estate market analysis tool. This could occur if we encounter insurmountable challenges, data limitations, or if the models do not perform as expected, impacting the project's overall success.

Methodology

1. **Data Acquisition and Preprocessing**:

- Acquire a comprehensive dataset from *https://emirates.estate*, capturing features like property size, number of bedrooms, house size, number of bathrooms, type of property, and location quality.

- Collate and preprocess reviews, involving cleaning tasks such as removing punctuations, converting to lowercase, eliminating stop words, and performing lemmatization.

2. **Preprocessing of Structured Data**:

- Standardize or normalize numerical attributes, namely property size, number of bedrooms, bathrooms, and house size.

- Process the categorical 'location' feature via methods like one-hot encoding or target encoding.

3. **Transformation of Unstructured Data**:

- Utilize a pre-trained NLP model, possibly a transformer model like BERT or GPT-3, to transform cleaned text data (reviews) into embeddings. This will convert each review into a semantically-rich high-dimensional vector.

- Extract these embeddings to represent the unstructured textual data for each property listing.

4. **Integration of Structured and Unstructured Data**:

- Concatenate the numerical vectors (which signify the reviews) with the structured dataset (like property size, number of bedrooms, etc.) to form a unified feature set for each property.

5. **Neural Network Model Training**:

- Implement and train a neural network-based regression model that predicts property prices using the combined feature set (incorporating both structured and unstructured data).

- Carefully tune the model considering the computational demands and complexity of the data. This may require iterative training and adjustments for optimal performance.

6. **Deployment and Application Development**:

- Design and build a user-friendly web-based application using Django or Flask.
- Embed data visualization tools (e.g., Matplotlib) for effectively displaying analysis results and property price predictions.
- Leverage SQLite for systematic database management, facilitating efficient storage and retrieval of real estate data.

7. **Continuous Model Updates and Retraining**:

- Utilize Google Colab for training and periodically retraining of the neural network model, ensuring its predictions align with the evolving Dubai real estate market of 2023 and beyond.

8. **Feedback Integration and Iterative Improvements**:

- Integrate a feedback mechanism within the website for users (real estate agents, buyers) to share insights, which will aid in refining the models and website functionalities.

9. **Documentation, Version Control, and Collaboration**:

- Maintain detailed documentation outlining the development, models, and insights generated.
- Implement Git and GitHub for version control, streamlining collaboration and development progress tracking.

Through these methodologies, the application aspires to stand out as a vital tool for Dubai's real estate stakeholders, empowering them with in-depth analysis drawn from both structured and unstructured data, ultimately facilitating astute property investment decisions.

Project plan

Required Work	Deadline
Bibliography and check for equivalent software (Gather information)	Week 3-4
Understand the data, its sources, and generation process	Week 6-7
Elicitation and Requirements of the functionalities of the software	
Choosing Machine Learning Models	Week 10-11
Prototyping & Presentation	Week 15

Team members

Samia Mahdaoui:

I am a senior in computer science at Saint Louis University. I have a very big passion for mathematics and this has led me to discover its incredible use in fields such as machine learning and AI, which is very exciting to me. During my summer, I was able to work on Supervised ML and Neural Network certifications which have enabled me to get more insight about the world of AI and push my understanding forward. This consequently led to me working on various ML projects using libraries such as Numpy, Tensorflow, and scikit learn. I have many hobbies, such as bodybuilding as it teaches me discipline and that there is always a way to improve and this is also true in life. Climbing is also one of my favorite activities because not only is it challenging physically but you get to analyze the obstacle courses to find the best path to the top.

Joe Farah:

I am a senior Computer Science student at Saint Louis University - Madrid Campus. My experiences include serving as a Teacher Assistant, a Python Tutor specializing in Object-Oriented Programming, and a Java Tutor focusing on Object-Oriented Software Design – all at Saint Louis University. I also showcased my web development skills during a stint with Wunderman Thompson and delved deep into the domain of Microsoft Dynamics 365 Finance and Operations as a Developer at Info-Sys. My technical proficiencies span MVC, Visual Studio, Java, Python, and machine learning tools like NumPy, Scikit-Learn, and TensorFlow. Beyond the technical realm, I passionately follow Formula 1, eagerly supporting Ferrari, and am an enthusiastic fan of football, with Manchester United being my team of choice. Another hobby of mine is playing chess. Chess is not just a game but a mental exercise that stimulates cognitive skills, strategic thinking, and patience. It's a blend of logic and creativity, similar to programming, making it an enriching pastime for someone in my field.

References

Bykovskii, G., & Dutot, V. (2022). Machine learning and artificial intelligence in a real estate marketing: a systematic review. *ResearchGate*. https://www.researchgate.net/publication/365781052_Machine_learning_and_artificial_intelligence_in_a_real_estate_marketing_a_systematic_review

Montgomery, M. (2022). How To Predict Real Estate Prices | Rev Real Estate School. *Rev Real Estate School: Canadian Real Estate Agent Coaching*. https://www.revrealestateschool.com/tips/how-to-predict-real-estate-prices