redia Real estate investment guide

Rana Akel Sachin Chandrasekhar Mandeep Jammu

Apurva Reddy Kistampally Kai Tai Tam Akhil Tandon

1. Introduction

We believe there is a market opportunity to develop a platform that will help experienced and new, individual and institutional investors identify properties ideal for investment to achieve maximum gain for their chosen strategy. We propose building a cloud based platform that will visualize ideal properties on a map of a chosen region. (Lu et. al, 2013). However, we will not use Lu's tool because the API 's are based on SQL, which we are not planning to use when extracting the data.

Identifying investment properties is a time-consuming and challenging task due to the lack of aggregated data availability. In this project we propose a new platform "redia", which will aggregate the data, analyze it for profit or cash flow margins and then visualize all the available properties on a map. (Benjamin, Sirmans & Zietz, 2001) Investors tend to invest in real estate compared to other financial assets because of the return distributions. In this project, we are going to use different return criteria based on their strategy of investment such as cash flow and growth.

2. Problem Definition

2.1. Current market and practice

Currently the real estate market primarily consists of individual investors who identify properties manually or large corporations who have dedicated departments. There is gap in the market for a platform that can use publicly available data and analyze it to identify ideal investments. A few startups that have identified this gap and provide individual listings that make good investments, however they are only available in few large cities and have not scaled up.

2.2. Innovations

2.2.1. Visualization - We are presenting the data in investor centric way, which no other platform in the market does. Most platforms like Zillow are homeowner centric, who have very different motivations

2.2.2. Analysis - We are applying neural networks to rental price prediction, which is a method that fairly novel to this use case. We expect that neural networks will perform better than traditional machine learning approaches due to their ability to determine interaction terms.

2.2.3. Data collection - We have collected tax information by integrated into county websites for accurate data. There are 3000 counties in the US, and each one requires customized integration.

3. Survey

3.1. Our approach and differentiation

We would like to solve the gaps identified above by providing a platform that simplifies the grunt work of identifying properties that offer the best investment opportunities. (Greene, 2019) Chapter 3 describes the manual steps real estate investors have to follow to find a rental property to invest in. To improve, we would automate the mathematical or numerical aspects of identifying a good deal.

Chandangope and Pragnesh visualized the housing prices all over US considering the forecasts for next 1 year from the Zillow website. (Chandangope and Pragnesh) We plan to integrate with a host of different data sources and aggregate it instead of relying on one source, as there are many factors to consider while evaluating the values. BAILEY, J. (1984) Proximity to colleges and hospitals are two of the variables that affect housing prices (Rivas et al, 2019). Our improvement to this would be to add visualization and automatic metric calculation on all the properties viewable on the map. Spacial statistics can be used to visualize real estate rental values. (Schernthanner, H., Asche, H., 2016) We agree that rental properties

should be determined based on finely tuned spatial distributions, thus we can use the formulae determined by these researchers during our visualization phase.

The prices of the houses were estimated using the data from 1980 through 2011 based on the supply and demand in the market. (Sheridan Titman, Ko Wang, Jing Yang) We may not use the main idea described in this paper to estimate the housing price as it mainly relies on the correlation between market supply and demand which might fail sometimes. The risk/reward can differ in each country based on local environment. (Geurts & Jaffe, 1996) We will only focus on US market and therefore only look at the institutional factors in the US. Housing can be effected by regional and local variables in the US. (Serena Ng & Emanuel Moench, 2010). We will use these factors to properly estimate the valuations of real estate and rent in the properties. Significant investment is being made in technologies to disrupt real estate market, such as IoT, Big Data, AI, SaaS and AR/VR. (Ullah, F., Sepasgozar, S., & Wang, C. 2018) We will incorporate few of these technologies in our platform to gather and estimate valuation.

3.2. Our customers

The audience is primarily the real estate investors, both institutional and private. There are a lot of platforms that help homeowners identify and estimate homes based on criteria, but not many that help real estate investors. Ross and Zisler discussed that the risk of investing in real estate is overstated. They reduced the variance of the risk of investing in real estate by a significant range, indicating investing in real estate is a reasonable method for generating profits. (Ross & Zisler, 1991). Therefore, smaller investors who do not have the means to accumulate and analyze the data can also maximize their gains using this tool. This will also allow individuals who previously have thought about investing in real estate, but are held back due to the lack of knowledge and intensive analysis.

Dickerson, A. M. (2009) describes how much power government has over policies of home buying which could encourage or discourage home ownership. Our approach is to show users most recent data and tax rates set by the IRS in order to make their decision based on current facts and real data.

3.3. Success Criteria and Impact

The success of the platform will be based on how well it does on identifying the best properties for investment. We can measure the accuracy of the platform by comparing the estimates to the ground truth metrics calculated manually. People expect house prices to be affected by their recent experience. (Case, Karl, John, and Robert, 2003) We would like to encourage people to invest in real estate based on current market conditions and tax rates as opposed to being based on historical data and recent experience. This platform would help investors and owner occupant buyers by helping them make data driven decisions. French studied how decision theory plays a role in real estate investment. He concluded that there is a difference between what customers say they would do and what they actually did. Market sentiment has an important effect in the real estate market (French, 2001).

3.4. Risks and Payoffs

The biggest risk is availability of data, and effort required to process and normalize it. In order to estimate the value of a property for investment, we would need to either estimate or find the property value, rent, financing costs, maintenance costs and taxes. (Kaufmann, V., & Wirsing, M. 2011) The estimations for value and rent can depend on many factors, eg school district, condition of the property, proximity to certain businesses. (Turner, 2015) Examples include determining cash flow and cash on cash ROI. This platform could exploit a previously untapped market of potential real estate investors that found it difficult to identify which properties to invest in.

4. Proposed Method

4.1. Data collection

The data collection part started by getting all the zip codes for Travis county in Texas. Smartystreets API was used to collect all the zip codes and store them locally. The next thing to do was to get all the homes for sale for each of the zip codes. We explored several APIs and websites. Some of the APIs were paid and some of them needed an approval process. In the meantime, we decided to get the data from Zillow and web scraping. We wrote a script to generate the URLs from the homes sales for each of the zip codes previously saved. Then we got up to 25 listings per zip code and saved the html pages. After that, using

some scripting and html parsing, we got the detailed url for each of the listings and saved the html pages locally. This was done partly mainly using Google Chrome extension and selenium python library for opening web pages. After having the html locally, we used Python and BeautifulSoup to parse the fields and got the required fields that we want to display on the map in our visualization. Once we got access to the Bridge Interactive API, we used the API to get the data in json format. There were limitations to the number of records we could retrieve and the offset used. As a workaround, we used the coordinations of the zip codes for Travis county longitudes and latitudes, collected earlier, and passed them to the API calls. That way, we were to collect around 110,000 records and pass them through our data wrangling and cleaning part. Since the data was made available for free for this class project, the property prices were rounded off to the nearest 100,000, and the resolution of images were reduced. We also used another data source called RentalBeast to get more rental data (6377 rows). They had an easy to use API which gave us the data we needed in only one API call. The data was also already clean, and ready to be used for modeling.

4.2. Infrastructure setup

For this Project after exploring multiple options, we have chosen to set up the Infrastructure on AWS. The main reason for choosing AWS was low cost and flexibility. As part ofAll the unstructured HTML data for each listing which has been collected from the Data Collection Phases will be stored in EC2 instance in AWS. After that using python scripts we parse the HTML data and create csv files which will be stored in EC2 instance.Once we have the data in a structured format(csv) the data is loaded into a PostgreSQL database which is hosted in Amazon RDS.As part of this project, we created 3 containers ,one for the Front End code, one for the Back End code and one for the Data layer which includes the database as well.We are using Docker to create different containers and then deploy them on AWS ECS.We have also implemented setting up of Database in AWS RDS and loading the data from extracted csv files into PostgreSQL database which can then be access by the application. Dockerization of the code base into 3 different containers is performed which will help in setting up the whole infrastructure either locally or even to a cloud service using a simple command which is provided in the Readme file.Using Docker compose, we were able to quickly deploy and scale applications into any environment both locally as well as hosting them on cloud.Running Docker on AWS provides a highly reliable, low-cost way to build, ship, and run distributed applications at any scale.

4.3. Data wrangling/feature extraction

For home sale data, there were over 110,000 records in the history of home sales in Austin. Each record had over 200 columns that contained information such as bathrooms total, listing price, and annual tax amount. In the data cleaning phase, we decided to only select the latest record, if the identical house was listed more than once. All the duplicates were also removed. Additionally, all the properties that had missing values in bathrooms, bedrooms, listing price, or square feet were also removed from the dataset. After all the cleaning had been done, there were approximately 47,000 records remaining. A few new columns were also generated for our conveniences, such as latitude, longitude, and IDs. A total of 25 columns were extracted from the data.

In addition to home sale data, we also collected rental data from Zillow in order to predict rental value using machine learning models. This data needed to be cleaned and converted to specific formats for model development. For example, removing thousand separators as well as a dollar sign before each price. There were also some inconsistencies in the data, such as showing the rental price or monthly mortgage payment in an incorrect column.

4.4. Model development

We have completed research into various machine learning approaches that could be applied to our rent prediction use case and have narrowed down to two major approaches. The first and preferred approach is to construct a neural network to model rent prices based on factors including number of bedrooms, bathrooms, square footage, age, and location (defined by latitude and longitude coordinates). We propose applying a neural network to this use case for two reasons. Firstly, the neural network is expected to have much stronger performance. For example, neural networks can grasp the interaction between input variables, regardless of whether they are independent or dependent. Secondly, we consider the application of a neural network to this use case to be a particularly novel task. As a fallback option, we will train a

few traditional machine learning models, such K-Nearest Neighbors, gradient boosting, and random forests. K-nearest-neighbors in particular has been historically shown to perform well on geographically distributed data such as latitude and longitude. Our models will have one additional preprocessing step to normalize the variables before feeding them to the algorithms, using the StandardScaler preprocessing function available in scikit-learn.

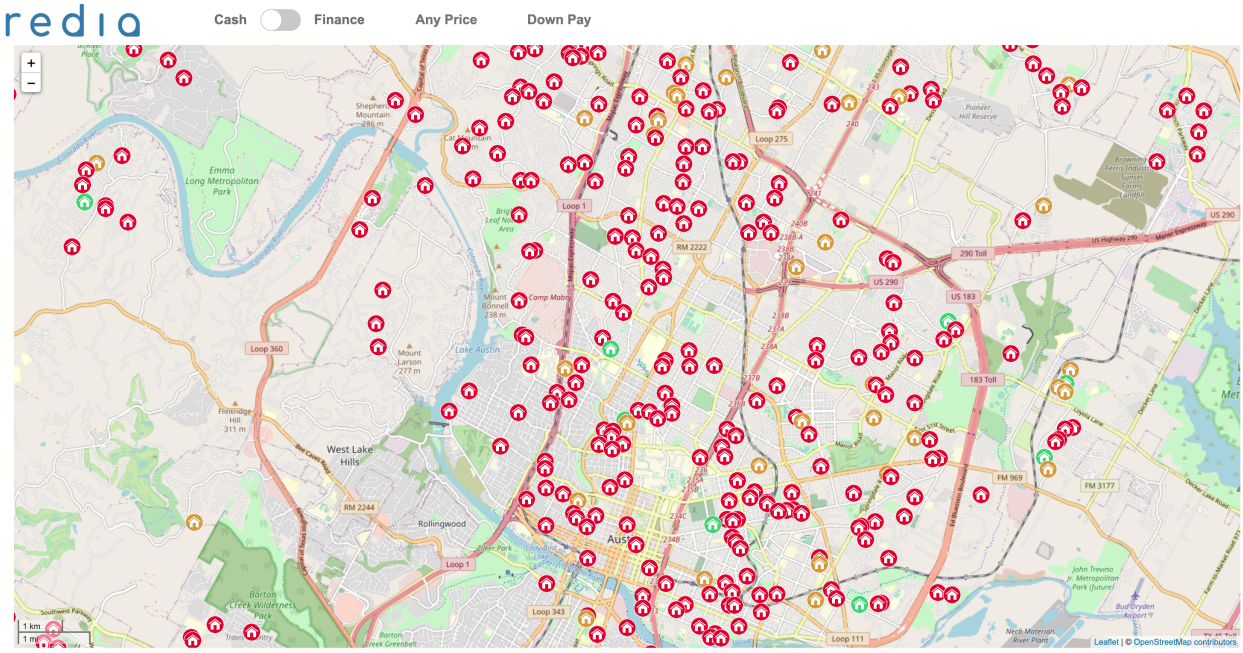
4.5. Backend development

Entire API for the backend is proposed to be done using Python. A YAML file generated from the UI is used to determine all the UI calls/actions needed. This YAML file is fed into the Swagger.io editor tool which is used for generating the required stubs for Python Flask. Once the stubs are generated, the UI function calls are filled in to return the appropriate values. The PostGresSQL created is connected here through Python using psycopg2 to read the required data including details of the property and also the rental value estimated using machine learning model which is saved in the database.

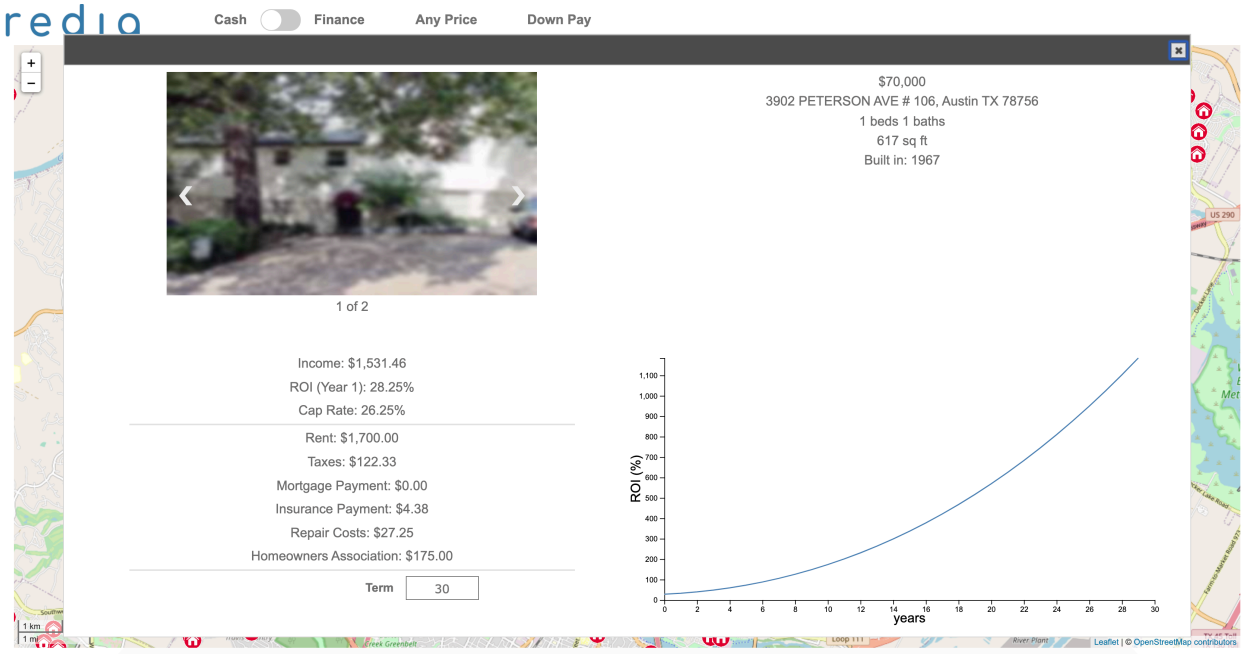
After having the data read, using the filter values from the UI like cashDown, interest rate, term etc., a score is generated to determine how profitable a property is. This score is then passed to UI which further helps in color coding the properties. Along with the score, fields like returnCash, returnMortgage, repair, insurance were also calculated appropriately and passed to UI for display in order to ease user's understanding.

4.6. Frontend development/Visualization

The base of the visualization is a map with markers representing each property. The markers will be color coded based on the recommendation level of the analysis. Hovering over the markers will reveal a popup



containing basic information such as the price, and clicking on the markers will open up a new window which will display detailed information about the property, including images, number bed/baths, area, etc. as well as the analysis that was used to recommend the property, including cap rate, ROI, income from rent, expenses etc. The visualization uses javascript/JQuery for basic DOM manipulation and visualization, leaflet library to render the interactive map and markers in the correct coordinates, OpenStreetMap for map layers, Fetch library to interact with backend using RESTful API's, and D3.js to visualize the data on a chart.



5. Experiments

5.1. Which features are most relevant to rental-price prediction model?

Our initial plan was to use industry standard feature selection techniques, such as cycling through different features combinations. We would measure the performance of the model using an evaluation function such as mean absolute error to determine most important features. While we did perform this experiment, we were unable to use all the recommended features, because many of these features were either not available or distributed very differently in the inference dataset (the for-sale dataset). As a result, we selected only those features which were commonly available in between the rental (training) and for-sale (inference) datasets. The final features selected were bedrooms, bathrooms, square footage, latitude, and longitude.

5.2. Does the visualization make the decision process of investing in a property easier for the user?

We ran a user study, for a subjective feedback from the users on their ability to make quick decisions on property investment. We deployed the website on AWS and recruited our friends and family for feedback. We received valuable feedback from user study using convenience sampling method. We realize that our users are satisfied regarding how we show the data on the map. It provided an easy and quick way to identify properties that were good investments. Not only does it provide spatial information regarding where the properties are, but it also shows whether properties that are clustered together are more worthy to be invested. According to our analysis, where best properties are shown as green, we can see that those green properties are scattered all around Austin. It does not seem that one area is more worthy to invest than the other. As a result, we should say we made a correct and successful choice in how we visualize our analysis, because it does help our users process which properties are worth to invest and which aren't.

5.3. How should hyper-parameters be tuned for the models and the best model be selected?

We used a mixture of grid-search cross-validation and randomized search cross-validation on our traditional machine learning models. On the neural network, we manually tested various neural network architects (activation functions, number of layers, and number of hidden units per layer). We used grid-search cross validation for the K-Nearest-neighbors model, varying the number of neighbors, the weights (uniform vs distance), and distance measure (euclidean vs Manhattan taxicab). For our tree based models, we tuned the regular parameters (number of estimators, max features, max depth, etc.) as a part of a randomized search cross-validation. To select the best model, we applied the best estimator output by

each cross validation training session, and predicted on completely unseen data. We expected the neural network to perform the best on this testing data, however the random forest performed the best, by a very large margin (mean absolute deviation of 84 vs 267 for the neural network). Upon further investigation, we discovered that pretty much every variation in architecture of the neural network was resulting in the same estimator. We likely did not have enough data, with only 6377 rental records, for the neural network to perform well. So we decided to keep the random forest model for our final product. When applying the random forest rental prediction model to the for-sale data, we had mostly reasonable results, with a small selection of properties that were producing anomalous rental values.

5.4. What is the minimal latency for the data to be visualized on the platform?

We wrote python scripts and open source tools to measure response time for data to load on the map. Populating the map with the appropriate icons was the worst offender in terms of latency. Our next step to improve this latency would be to paginate the API and populate the map progressively, rather than in one shot.

5.5. How do we collect the data that is necessary for the project?

We looked at the data from various sources including Zillow, Bridge Interactive, Austin, TX county property tax website, to find best possible data for analysis. It turned out that Bridge Interactive API returns the most useful results, as most of the home sale properties were extracted from Bridge Interactive API. The amount of data was big enough that even though we eliminated a massive part of data due to various reasons, such as incorrect data and missing values, the remaining dataset was still sufficient for us to generate machine learning models and finish the analysis.

6. Conclusions and discussion

We have successfully created a platform, Redia, that can potentially help both experienced, new, individual, or institutional investors identify properties ideal for investment to achieve maximum gain for their chosen strategy. This cloud-based platform is able to visualize ideal properties on a map of a chosen region. For this project, we only started with a small city in Texas, Austin. However, with the way we set up, it has the potential to scale up to the US.

Redia should help real estate investors analyze their potential profit for investing in different properties.

In the future, additional detailed analysis can be imported to the analysis page. We will also need to receive feedback from more users in order for us to decide which feature users think is best to help them decide whether a property is worth investing in or not. In terms of the visualization, we find that the properties that are being assigned anomalous rental values have an unduly influence on the score, and thus the coloration of the icons. It would be worth it to discover the root cause of these anomalies, and fix it.

All team members have contributed a similar amount of effort.

Reference

Benjamin, J., Sirmans, S., & Zietz, E. (2001). Returns and risk on real estate and other investments: more evidence. *Journal of Real Estate Portfolio Management, 7*(3), 183-214.

Case, Karl E., John M. Quigley, and Robert J. Shiller. (2003) Home-buyers, Housing and the Macroeconomy. (2003). *RBA Annual Conference Volume (Discontinued)*

Dickerson, A. M. (2009). The myth of home ownership and why home ownership is not always a good thing. *Ind. LJ*, *84*, 189.

French, N. (2001). Decision theory and real estate investment: an analysis of the decision-making processes of real estate investment fund managers. *Managerial and Decision Economics, 22*(7), 399–410. doi: 10.1002/mde.1029

Greene, D. (2019). Chapter Three - How to Find Deals. *Buy, rehab, rent, refinance, repeat: The BRRRR rental property investment strategy made simple.*

Kaufmann, V., & Wirsing, M. (2011) Real Estate Valuation and Investment.

BAILEY, J. (1984), "REAL ESTATE INVESTMENT ANALYSIS", Journal of Valuation, Vol. 2 No. 4, pp. 356-365. doi: 10.1108/eb007959

Tom G. Geurts & Austin J. Jaffe, 1996. "Risk and Real Estate Investment: An International Perspective," Journal of Real Estate Research, American Real Estate Society, vol. 11(2), pages 117-130.

Li, M., Bao, Z., Sellis, T., Yan, S., & Zhang, R. (2018). HomeSeeker: A visual analytics system of real estate data. *Journal of Visual Languages & Computing, 45,* 1–16. doi: 10.1016/j.jvlc.2018.02.001

Lu, Y., Zhang, M., Li, T., Guang, Y., & Rishe, N. (2013). Online spatial data analysis and visualization system. *Proceedings of the ACM SIGKDD Workshop on Interactive Data Exploration and Analytics - IDEA 13.* doi: 10.1145/2501511.2501522

Moench, E., & Ng, S. (2011). A hierarchical factor analysis of U.S. housing market dynamics. *The Econometrics Journal, 14*(1). doi: 10.1111/j.1368-423x.2010.00319.x

Rivas, R., Patil, D., Hristidis, V., Barr, J. R., & Srinivasan, N. (2019). The impact of colleges and hospitals to local real estate markets. *Journal of Big Data, 6*(1). doi: 10.1186/s40537-019-0174-7

Ross, S., & Zisler, R. (1991). Risk and return in real estate. *The Journal of Real Estate Finance and Economics, 4(2).* doi: 10.1007/bf00173123

Schernthanner, H., Asche, H., Gonschorek, J., & Scheele, L. (2016). Spatial modeling and geovisualization of rental prices for real estate portals. In O. Gervasi, B. Murgante, S. Misra, A. M. A. C. Rocha, C. M. Torre, D. Taniar, B. O. Apduhan, E. Stankova, & S. Wang (Eds.), *Computational Science and Its Applications—ICCSA 2016* (Vol. 9788, pp. 120–133). Springer International Publishing. https://doi.org/10.1007/978-3-319-42111-7_11

Turner, B. (2015). Chapter Five - Analyzing a Rental Property. *The Book on Rental Property Investing: How to Create Wealth and Passive Income Through Intelligent Buy & Hold Real Estate Investing!.*

Ullah, F., Sepasgozar, S., & Wang, C. (2018). A Systematic Review of Smart Real Estate Technology: Drivers of, and Barriers to, the Use of Digital Disruptive Technologies and Online Platforms. Sustainability, 10(9), 3142. doi: 10.3390/su10093142

Rivas, Ryan & Patil, Dinesh & Hristidis, Vagelis & Barr, Joseph & Srinivasan, Nani. (2019). The impact of colleges and hospitals to local real estate markets. Journal of Big Data. 6. 10.1186/s40537-019-0174-7.

Serena Ng & Emanuel Moench. (2010). A Hierarchical Factor Analysis of US Housing Market Dynamics. Columbia University