Rivalry Beyond Campus: Identifying Real Estate Opportunities for School Districts using Machine Learning

Yiheng An (805640602), Shuangfei Li (705636907), Jiaxing Wang (305645225), Zhuoyue Ran (005867034)

***Abstract*— In view of the increased uncertainty in the real estate market over the pandemic of COVID-19, it is more important for investors and customers to avoid uninformed decisions. In this project, we built a machine learning framework to estimate the expected value of residential properties in the university district of both UCLA and USC. We developed a set of regressors over tax assessment data and Zillow housing data and identified 5 investment opportunities for each school district. We also explained such results based on a naïve value discovery mechanism. The Stacked Ensemble model and SVR model in our system showed the best out-of-sample performance in terms of MAE and MSE respectively. The district of UCLA is a better submarket because of its higher opportunity density.**

***Index Terms*—Machine Learning, Value Discovery, Real Estate**

I. Introduction

Due to the impact of COVID-19, the real estate market has been exposed to increasing uncertainties with more complex relationships between asset value and a huge amount of factors. From a macroeconomic perspective, these factors may include market sentiment and confidence, demographic features of regions, monetary policies, financial regulations, public health issues, and so forth, some of which are not easy to be controlled or even observed.

Fortunately, the innovation of machine learning algorithms and the revolution of computing power have provided new possibilities for value discovery. These techniques allow people to better understand the price variability in the real estate market and unleash the potential of machine-based asset valuation. In this project, we will investigate two geographical submarkets in Los Angeles County.

A. Problem Statement

To be specific, we would study the residential properties in the district of two universities, UCLA and USC, and try to answer two research questions:

- What are the major drivers for the value of properties?
- Can we identify 5 opportunities for each district?

To approach this problem, we would formulate such two questions as a machine learning problem and develop a solution that includes one or more robust ML models with good potential of generalization and an appropriate mechanism of interpretation to adapt to different geographical segments. Apart from that, we are also interested in which university district is overall better in terms of real estate investment.

B. Task Description

In this project, we are going to deal with a regression-based pricing problem through a series of machine learning methods. The goal is essentially to identify residential real estate opportunities in the district of UCLA and USC. So, the first task is to set up a set of reasonable rules to answer two fundamental questions:

- What is the research scope of a school district?
- What is the opportunity discovery model? (i.e. How a property can be theoretically seen as an opportunity?)

Besides, we need to translate our opportunity discovery model into a doable data science pipeline, which includes a data streaming process and supervised learning setup. We would evaluate out-of-sample performance and explore the key drivers of the property price for each school district.

Moreover, based on the performance of the best algorithm, we hope to identify at least 5 investment opportunities of residential properties for both the UCLA district and the USC district, and answer which school district is the winner overall.

II. Statement of Previous work

This section covers some previous research papers and projects that inspired our work. Vahid Moosavi[1] built an innovative machine learning portal based on open-source data to estimate the rental and sale price indexes in Switzerland. He found that Random Forest was able to acquire reasonable results with the median absolute relative error of 6.57%.

Dr. Swapna Borde and other team members[2] compared the performance of Linear Regression, K-Nearest Neighbor Regression and Random Forest Regression in predicting real estate price trends. They found that the Random Forest model obtained the smallest errors in terms of RMSE, MAE and MAPE.

Alejandro Baldominos and his co-workers[3] developed an application of diverse machine learning techniques, including Ensemble Method, Support Vector Regression and Multi-layer Perceptions with the objective of identifying real estate opportunities for investment. Their results indicate that outperforming models are always those consisting of Ensembles of Regression Trees based on the cross-validation performance.

Shashi Bhushan Jha and his team[4] trained several machine learning classifiers over a ten-year actual dataset to predict whether the actual sale price of a property is greater than or lower than the appraised price of the property. Based on the evidence of Logistic Regression, Random Forest, XGBoost and so on. They reached the conclusion that XGBoost is the best algorithm to deliver accurate prediction results.

Dieudonné Tchuente and Serge Nyawa[5] experimented with seven machine learning techniques upon 5 years of transactions data in major French cities to estimate real estate prices. Taking advantage of a high level of data granularity, they clearly observed that there were very important differences regarding

the forecasting errors between high-cost cities (e.g., Paris, Bordeaux, and Nice) and medium-cost cities (e.g., Toulouse, Lille, and Montpellier). They believed it would be more relevant to train specific models for some geographical submarkets (cities in their case) rather than global models including all cities.

Based on these prior works and what we learned from this course, together with our time constraints, we will select a series of techniques such as K-Nearest Neighbors, XGBoost, Random Forest, Support Vector Regression and Stacked Ensembles with the help of AutoML techniques to create our ML system and evaluate its performance over out-of-sample results.

TABLE I
SELECTED FEATURES AND TARGET

Name	Description
Location Features	
zip2	The 5-digit zip code that matches property's actual street address
distance_to_ucla	The miles distance away from UCLA
distance_to_usc	The miles distance away from USC
PropertyLocation	The actual address of the property (used to match real-time listing price)
Property Characteristics Features	
Bedrooms	The total number of bedrooms
Bathroom_per_bedroom	The total number of bathrooms/ total number of bedrooms
Units	The total number of living units
YearBuilt	The year property was originally built
Years_until_effective	The number of years between build year and effective year
Price Features/Target	
LandValue_percent	The proportion of a property's land value to its total value
Price_per_unit	The total value/ total number of living units
ZHVI	Zillow Home Value Index, the typical (35th to 65th percentile range) home value of a zip code
ZHVI_sf	The price per square foot for a typical property: Zillow Home Value Index / the total square footage
price_sf	The price per square foot for a property (the target of models)
…	…

III. Major Challenges and Solutions

Based on the structural nature and expected functions of our predictive framework, there are at least two major challenges that should be understood properly: the opportunity discovery mechanism and the data proxy issue.

A. Opportunity Discovery Mechanism

Researchers have developed many financial models with different flavors and structures to evaluate residential properties, but few of them integrate well with our machine learning methods due to the unattainable granularity of financial data. So, here we try to capture some basic ideas of property evaluation theories and formulate them into a machine learning approach.

We believe that the price of a property would deviate from its expected value due to some uncaptured information and unpredictable factors[3]. We use p_i and $\bar{p}_i$ to denote the market price and expected value of property i. The property would be identified as an investment opportunity when $p_i < \bar{p}_i$, because it could generate immediate profit if sold right after its purchase, assuming no significant transaction costs. Then, we would be able to train a machine learning model $f: \mathbb{R}^d \rightarrow \hat{p}_i$ over d dimensional data to estimate $\bar{p}_i$ for each in-sample property. And if $p_i < \hat{p}_i$, the model predicts property i is an opportunity.

B. Proxy Selection

Based on the opportunity discovery model proposed above, the next challenge is collecting relevant data to make model estimations reasonable. As we know, for property i , its expected value $\bar{p}_i$ is always unrevealed. Feeding the model with proper proxies is important to the model success. Also, although market price p_i is always available on paper, the majority of records are associated with closed deals. So, these backward-looking signals would introduce the risk of bad generalization, thus making our application unpractical.

To address these problems, we choose the tax-based assessor parcels data as our major dataset to estimate the expected value of properties. We believe it would contribute as a good proxy because unreasonable evaluations directly lead to tax loss or crisis of social justice. Besides, we turn to Zillow data to make sense of the market as well. To be specific, we use prices of historical deals to adjust the estimation of $\bar{p}_i$, and choose the listing Zillow estimated prices of properties as the proxy of p_i to make our predictions forward-looking.

IV. Data Acquisition and Description

The datasets used in this project can be roughly divided into three categories: location data, property characteristics data, and price data. The location data and property characteristics data are obtained from Assessor Parcels Data as part of County of Los Angeles open data, the price data is collected from the same dataset and extra Zillow housing data.

The raw datasets include 38 million properties with descriptions for parcels on the assessor's annual secured assessment rolls from 2006 to 2021. In order to have a good consistency between datasets, we filtered properties by restricting their assessment roll year between 2013 and 2021. Selected features are summarized as Table I.

V. Workflow and Analytic Techniques

Our workflow design can be divided into four main components. They are data preparation, intermediate feature processing, predictive modelling and model evaluation. This structure of thinking can be summarized as the experimental diagram shown in Fig. 1.

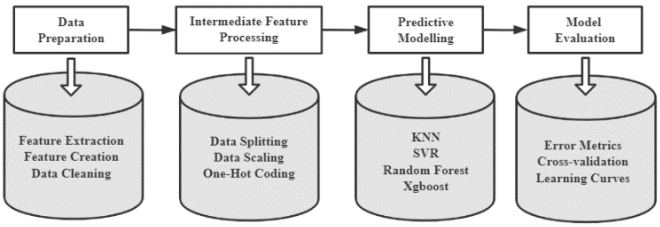

Fig. 1. Workflow design diagram

A. Data Preparation

Feature Extraction

Location features include geographical information that is related to each property. In this project, we will use zip code, address, longitude and latitude as necessary location identifiers. Based on these features, we filtered and retained the properties within 3 miles away from UCLA and USC as our study scope visualized in Fig. 2.

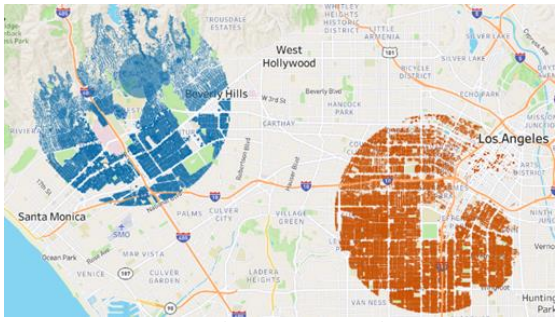

Fig. 2. The study scope of school districts of UCLA and USC

As for characteristics-based features, the selection is based on both the physical nature and the temporal nature of properties. We choose a series of features such as the square footage, number of units, number of bedrooms, and so on to reflect the physical nature of each property. Also, we retain built year, effective year, and so on to capture the time-based nature of properties and facilitate feature engineering.

For the price-based data, we chose the total value and its detailed components for each property to reflect the assessor's expectation; we introduce Zillow real-time data and Zillow Home Value Index (ZHVI), a zip code's typical home value, to capture market expectation. Finally, we use Zillow real-time price data as a necessary benchmark when interpreting model outputs.

Feature Creation

We calculated the property price per square foot as the target of potential regression models and created a series of proportion-based features and difference-based features, such as the land value proportion, number of bathrooms per bedroom, the number of years between build year and effective year, and so on. These features allow us to keep the information as much as possible if we have to drop some features to mitigate multicollinearity and potential confounding.

Also, we matched each property with the ZHVI by zip code and assessment date (January of each roll year), integrating the market price signal at that time. Based on this, we would be able to calculate the price of a typical property (35th to 65th percentile range) for each zip code discovered by the market at the valuation date, introducing market-based leading indicators to our expectation estimation.

Data Cleaning

After combining the target and all created features together, we first addressed the issue of missing values due to zero denominators. We imputed these observations with a special value to denote their uniqueness. Next, we corrected the data type for each feature and renamed them for simplicity. Moreover, we used the correlation heatmap of numerical variables to identify potential multicollinearity, and decided to remove features with more than 0.95 correlation scores. Finally, we regarded records that fall outside of ±3 standard deviations as outliers, and replace them with the feature mean without outliers to finish up the data cleaning.

B. Intermediate Feature Processing

Data Scaling

Due to the requirement of distance-based regression algorithms such as KNN, SVR, and so on, we need to scale our features in a proper way. For numerical features, we use the z-standardization method to normalize their ranges. The normalization algorithm follows the formula:

$$Z = \frac{x-\mu}{\sigma} \quad (1)$$

In formula (1), x represents a specific value of the feature, μ is the mean value of the feature, σ is the standard deviation of the feature. By conducting this algorithm for each value, the feature will have a mean of 0 and a standard deviation of 1.

One-Hot Encoding

To facilitate our machine learning modelling, we would further convert all categorical features through one-hot encoding, which binarizes categorical levels and pulls them together as a multidimensional matrix from the original feature vector space[6]. If a categorical feature contains N levels, the same number of binary features will generate.

Unlike the method of label encoding, one-hot encoding allows us to avoid ranking-based bias from level digitalization if we assume all levels of the same categorical features are equally away from each other.

C. Predictive Algorithms

Relevant research has indicated several powerful machine learning algorithms widely used in regression problems of real

estate pricing. Here we choose five of them to build a set of competing regressors in our system.

K-Nearest Neighbors (KNN)

KNN is a non-parametric algorithm to perform regression or classification. This distance-based method always requires feature normalization or standardization. Given a positive integer *K* and one observation x_i, the algorithm will identify the *K* points that are closest to x_i. Then, in a regression setting, the predicted output can be generated by aggregating the selected *K* points, such as computing the average. The most important hyper-parameter of KNN is the value of integer *K*.

Random Forest (RF)

Random forests operate by constructing a multitude of decision trees in the training process and outputting a prediction of individual trees based on the parameters inputted. These large number of decision trees are built over bootstrapped samples. If a simple decision tree model is trained on B number of bootstrap samples, then the prediction of the RF, denote as f_{RF}, will be the average of individual predictions, denote as f^*, coming from these decision trees. That is:

$$\widehat{f_{RF}}(x) = \frac{1}{B}\sum_{b=1}^{B} f^{*b}(x) \quad (2)$$

In each split, a random set of features is selected as split candidates, but only one of these predictors would be used in the split. In this way, RF algorithm has the advantage of decorrelating predictors, thus reducing the variability of the model performance.

Support Vector Regression (SVR)

Following the principle of the Support Vector Machine, SVR aims to find a function that presents a margin of tolerance from the target values. It maps training examples to points and tries to maximize the width of the gap between the predicted and true outputs. To address the more common situation that nonlinear relationships between target values, a special kernel approach will be applied to enlarger the feature space.

C could be one of the most important hyper-parameters when applying SVR. In general, when *C* is relatively small, the tolerance of violations to the margin would also be small. So, it always leads to a narrow margin. As the *C* goes up, we will tolerate more violations to the margin, thus leading to a wider margin.

XGBoost (XGB)

XGB is a tree-based ensemble method optimized to perform efficient and flexible predictions within the framework of gradient boosting[7]. Unlike the traditional tree building process, this method builds sequential trees using a parallelized implementation, allowing us to achieve relatively low prediction errors with modest memory and runtime requirements. Another advantage is its excellent performance over complex and high-dimensional data.

Stacked Ensembles (SE)

The thinking of Stacked Ensemble here is to use multiple learning algorithms (weak regressors) we built already to build a model with better predictive performance (strong regressor). Indeed, the RF algorithm is an ensemble-based leaner. The method of stacked ensemble tries to find the optimal combination of a collection of prediction algorithms using a process called Stacking or Super Learner. H2O ai has automated most of the steps in this algorithm[8], so now it is much easier to implement it in data science projects.

D. Model Evaluation Metrics

Mean Square Error (MSE)

As one the most commonly used matric, MSE is the average of the squared difference between the predicted value and true value, which is computed as formula (3). Since its measure of error, the smaller the metric value the better.

$$MSE = \frac{1}{n}\sum_{i=0}^{n}(y_i - \hat{y}_i)^2 \quad (3)$$

Mean Absolute Error (MAE)

Similar to MSE, MAE is another error metric of regression models, which computes the average of the absolute difference between the predicted value and the true target as formula (4). Compare to MSE, MAE does not penalize large errors or outliers so much.

$$MAE = \frac{1}{n}\sum_{i=0}^{n}|y_i - \hat{y}_i| \quad (4)$$

Median Absolute Error (MedAE)

MedAE is the median of the distribution of differences between predicted and true values, which represents the error value that is larger than 50% of all the prediction errors. As shown in formula (5), its nature of absolute difference allows it to work robustly over outliers.

$$MedAE = median(|y_1 - \hat{y}_1|, \dots, |y_n - \hat{y}_n|) \quad (5)$$

VI. EXPERIMENTAL IMPLEMENTATION AND RESULTS

A. Experimental Implementation

For each school district, we will follow an identical structure of experimental setting, but train models exclusively over the data of that district. This means, for each algorithm, we believe there should be different options of hyper-parameters for UCLA and USC. As shown in Fig. 3, for each predictive model, we will first conduct the hyper-parameters tunning based on 5-fold cross-validation over the training dataset (80%) to search the optimal parameters. Due to the cost of time and computation, here we only pick one or two key parameters for each model and hope to obtain a moderate option through this iteration process.

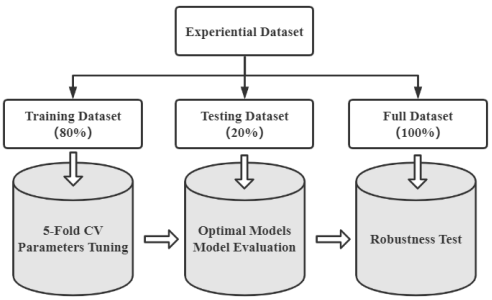

Fig. 3. Predictive modelling process

Given the fair option of hyper-parameters, we will test the out-of-sample performance for our models over the testing dataset (20%). Then, we would evaluate the robustness of our

best model by plotting its learning curve over the full dataset and estimate the expected value of testing set properties for two districts. Finally, we would be able to identify investment opportunities by comparing expected value and listing price.

B. *Out-of-sample Performance*

From Table II, we can see the testing set performance for each regressor over the properties of UCLA and USC district.

TABLE II
TESTING SET PERFORMANCE METRICS

Models	UCLA District			USC District		
	MSE	MAE	MedAE	MSE	MAE	MedA
KNN	0.963	0.645	0.509	0.985	0.707	0.548
RF	0.980	0.649	0.513	0.988	0.706	0.540
SVR	0.976	**0.630***	**0.483***	1.052	**0.663***	**0.421***
XGB	0.964	0.643	0.509	0.993	0.706	0.541
SE	**0.955***	0.639	0.506	**0.976***	0.702	0.540

In our framework, estimation errors described by MAE are lower overall compared to those explained by MSE. For both school districts, the SVR is the best model in terms of the smallest MAE and MedAE, and the SE is the best one in terms of the lowest MSE. For the UCLA district, RF is the weakest regressor among this model set in terms of all error metrics, while we would obtain different weakest regressors for the USC district based on each error metric.

C. *Robustness of Testing Results*

To further test the robustness of our best regression models, here we conducted 5-fold cross-validations for SVR models we learned before over different sizes of data. This means we iterated a similar process for our models over randomly selected experimental data from small to relatively large until the full we reach the full dataset. In this way, we can depict the learning curve and observe the consistency of our model performance.

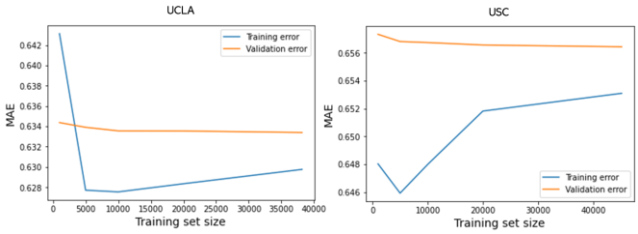

Fig. 4. The leaning curve for SVR models over UCLA and USC district

From the learning curve in Fig. 4, we can tell SVR has relatively stable validation error scores over various sizes of data for both university districts. This indicates the model is relatively robust and has a great potential of generalization. Also, we can note the training error score is almost lower than the validation error score for both districts, which is a signal of overfitting. As the size of the training set goes up, we can see two metrics became close to each other and the issue of overfitting was going to be mitigated.

D. *Feature Importance*

Feature importance analysis is a useful tool to open the black box of ML models and have a glimpse of the prediction mechanism. For many tree-based models, the importance score is always calculated by the amount that each split improves the performance measure, weighted by the number of observations the node is responsible for. Here we will use our XGB models to depict the global structure of predictors and answer our first research question.

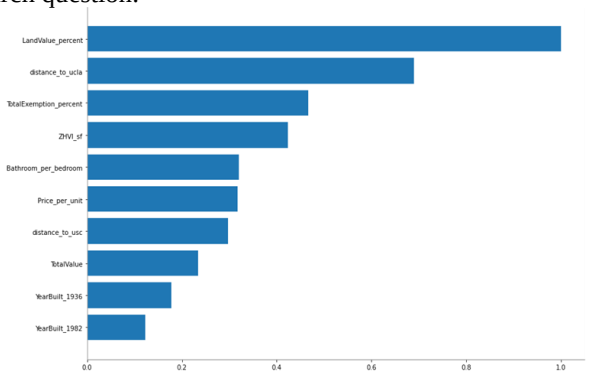

Fig. 7. Feature importance of XGB model for UCLA

In Fig. 7, the features here are listed from the most important ones to the less important ones. We can see that for the UCLA district, the land value proportion to the total value is the most important numerical driver, distance to UCLA is the second one.

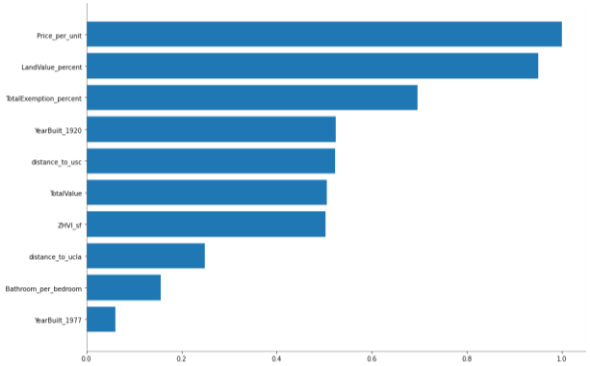

Fig. 8. Feature importance of XGB model for USC

For the USC district, as shown in Fig. 8, although land value proportion is relatively important as well, price per unit is the most important factor, distance to the campus is not as important as it is for UCLA. For both districts, the typical market price at the time of assessment we introduced from Zillow is one of the top 10 drivers.

E. *Investment Opportunities Discovery*

Back to our second purpose, here we identified 5 investment opportunities for each university district based on the suggestion of our best SVR models. We adjusted the estimation of the expected value we generated for each property through a difference-based multiplier and compare it to Zillow's Best Estimate, which represents the market's signal. If the expected value of one property is higher than the market's current pricing, it would be regarded as an undervalued property thus a potential opportunity.

TABLE III
OPPORTUNITIES DISCOVERY: MODEL EVIDENCE FOR UCLA ($/SQFT)

Undervalued Property Address	UCLA District	
	Expected Value	Zillow's Best Estimate
1821 S Bentley Ave APT 301, CA 90025	1432	1,028
1495 Roxbury Dr, CA 90035	1347	1,018
1444 Princeton St, CA 90404	1059	788
1447 Franklin St APT 4, CA 90404	1127	730
10375 Wilshire Blvd APT 7D, CA 90024	568	454

For the properties listed in Table III and Table IV, we believe they are undervalued and have the potential of obtaining a higher price per square foot in the future. From the perspective of pricing errors, the opportunities in the UCLA district promise better potential profits due to their larger pricing errors thus a stronger growth momentum. However, in terms of the absolute price level, the USC district still provides some quality opportunities with a lower investment threshold.

TABLE IV
OPPORTUNITIES DISCOVERY: MODEL EVIDENCE FOR USC ($/SQFT)

Undervalued Property Address	USC District	
	Expected Value	Zillow's Best Estimate
2271 W 23rd St, Los Angeles, CA 90018	496	436
2681 Orchard Ave, CA 90007	353	333
3932 La Salle Ave, CA 90062	499	423
4831 9th Ave, CA 90043	478	449
1150 E 41st St, CA 90011	395	373

In order to assess the overall investment opportunity for these two school districts, we also develop a naïve metric called opportunity density. For an out-of-sample dataset, the opportunity density for region *i* is calculated by its number of potential opportunities n_{oi} over the number of properties n_{pi} it has, which is identified as formula (6):

$$OD_i = \frac{n_{oi}}{n_{pi}} \times 100\% \quad (6)$$

Based on the 20% testing set, the opportunity density score of the UCLA district is 28.28% and the opportunity density score of the USC district is 22.25%. Therefore, we believe the UCLA district is a better submarket to invest in.

VII. SUMMARY

There are many imperfect solutions thus exciting opportunities of machine learning applications in real estate nowadays. This work is one attempt at ML-based property valuation and an exploration of regression solutions over geographic submarkets. In this project, we proposed a workflow to approach a price estimation problem for properties in 2 university districts. Many techniques in data cleaning, feature engineering and predictive modelling were involved. After identifying the best model, we applied the concept of feature importance to explore the key drivers of the expected property price. We also use the framework to propose 5 undervalued properties and assess the overall real estate investment attraction for both university districts.

A. Limitations

Estimating the expected value for properties in a reasonable manner is a complex and data-intensive problem. Subjecting to the limited time and computing power for data processing, some necessary steps such as feature selection and hyper-parameter tuning in our workload have been simplified. Also, due to the lack of financial data and industry immersion, we used a naïve method to adjust the bias between model output and the market signals, which implies that the estimations we made were probably not the optimal ones. Therefore, to further refine our framework, incorporating more industry best practice is important in the future.

B. Future Work

The ideas and solutions of ML-based asset pricing are quite transferable among any business which focuses on value discovery. In the future, we believe there are at least two aspects worth exploring.

Firstly, introducing more valuable features. Big data technologies have unleashed the great power of feature creation. Many novel features that contain valuable information remain to be coded and constructed. Besides, exploring the impact of time horizons. Although the best models in our system display good robustness, it is still not clear whether their performance is limited to specific of time windows. More specialized research in this aspect would be helpful to deepen our understanding of model adaptability and robustness in broader user scenarios.

VII. ACKNOWLEDGMENT

We would like to say thanks to those who had done previous research in the field of real estate pricing and evaluation. This project was inspired in part by their contributions. Also, we want to thank all Econ 445 teaching contributors, including Prof. Mendler, the TA, and guest speakers for their help and inspiration.

REFERENCES

[1] Moosavi V. Urban Data Streams and Machine Learning: A Case of Swiss Real Estate Market. *arXiv:170404979 [cs, q-fin, stat]*. Published online March 30, 2017. Accessed March 4, 2022. http://arxiv.org/abs/1704.04979

[2] Borde DS, Rane A, Shende G, Shetty S. Real Estate Investment Advising Using Machine Learning. Published 2017. Accessed March 4, 2022. https://www.semanticscholar.org/paper/Real-Estate-Investment-Advising-Using-Machine-Borde-Rane/627e72cfcf666922d8b59e1e014e6aea51acf69d

[3] Baldominos A, Blanco I, Moreno AJ, Iturrarte R, Bernárdez Ó, Afonso C. Identifying Real Estate Opportunities using Machine Learning. *Applied Sciences*. 2018;8(11):2321. doi:10.3390/app8112321

[4] Jha SB, Pandey V, Jha RK, Babiceanu R. *Machine Learning Approaches to Real Estate Market Prediction Problem: A Case Study*.; 2020.

[5] Tchuente D, Nyawa S. Real estate price estimation in French cities using geocoding and machine learning. *Ann Oper Res*. 2022;308(1):571-608. doi:10.1007/s10479-021-03932-5

[6] Yu, Lean, Rongtian Zhou, Rongda Chen, and Kin Keung Lai. "Missing data preprocessing in credit classification: One-hot encoding or imputation?." *Emerging Markets Finance and Trade* 58, no. 2 (2022): 472-482.

[7] XGBoost Documentation. https://xgboost.readthedocs.io/en/stable/

[8] Stacked Ensembles H2O 3.36.0.3 documentation https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/stacked-ensembles.html