DSE511: Introduction to Data Science and Computing (I)
Final Project- Part I
California Housing Prices

ALBINA JETYBAYEVA, The University of Tennessee, Knoxville, USA
PRAGYA KANDEL, The University of Tennessee, Knoxville, USA
ISIDORA FLETCHER, The University of Tennessee, Knoxville, USA
AMIREHSAN GHASEMI*, The University of Tennessee, Knoxville, USA

ACM Reference Format:
Albina Jetybayeva, Pragya Kandel, Isidora Fletcher, and Amirehsan Ghasemi. 2021. DSE511: Introduction to Data Science and Computing (I)
Final Project- Part I
California Housing Prices. 1, 1 (November 2021), 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The new computing technologies have widened the scope of machine learning to a great extent. It's ability to learn from previous computations and independently adapt to new data is making it popular across various disciplines. Various sectors such as business, bioinformatic, computer engineering, pharmaceuticals, medical, climate change, and statistics are using machine learning models to gather knowledge and predict future events [5]. One of the important sectors that machine learning can be used is on real estates to predict the prices of houses. Buying a new house is always a big decision. It gets affected by various factors such as location, size of house, quality of house, future trading price, school zone etc but prioritizing these factors is tough [4]. What would be more important? Is it the location or quality of the house? Machine learning can be used to ease the process of decision making by forecasting the house prices with maximum accuracy of the market trend and the building model based on historic data set [6]. What happened in the past and what was important, how it affected the price and what is going to happen? Prediction of house prices is not only limited to homeowners, but also equally important to real estate agents, appraisers, mortgage lenders, brokers, property developers as well as investors. The prediction of housing prices using ML is not a new concept. The selling price of houses of Pitt county, North Carolina was predicted by the use of parametric and semi parametric regression [2]. Bae and Park, 2015 used a machine learning algorithm for predicting the prices of houses in Fairfax county of

*All authors contributed equally to this research.

Authors' addresses: Albina Jetybayeva, The University of Tennessee, Knoxville, The Bredesen Center for Interdisciplinary Research and Graduate Education, Knoxville, TN, USA, 37996-3394; Pragya Kandel, The University of Tennessee, Knoxville, The Bredesen Center for Interdisciplinary Research and Graduate Education, Knoxville, TN, USA, 37996-3394; Isidora Fletcher, The University of Tennessee, Knoxville, The Bredesen Center for Interdisciplinary Research and Graduate Education, Knoxville, TN, USA, 37996-3394; Amirehsan Ghasemi, The University of Tennessee, Knoxville, The Bredesen Center for Interdisciplinary Research and Graduate Education, Knoxville, TN, USA, 37996-3394.

Manuscript submitted to ACM

Virginia. The ML approach to predict has also been used in various other countries such as Brazil. Afonso et al. 2019 [1], predicted the housing prices with deep learning and random forest ensemble in Brazil. In this project we are trying to understand the factors contributing to housing prices in California. We hypothesis that the values of the houses will be directly related to the number of total rooms, the number of total bedrooms and median income. And inversely related to the housing median age. We will also be investigating how the price is affected by the location (longitude, latitude and and ocean proximity), the population and the number of households.

2 DATA

We will be using the dataset for California Housing Prices: (https://www.kaggle.com/camnugent/california-housing-prices).

The dataset was used for the machine learning basics introduction in the book by Aurélien Géron 'Hands-On Machine learning with Scikit-Learn and TensorFlow' [3]. The data is chosen because it has an understandable list of variables and the optimal size between too small and big. The data contains information on houses from the 1990 California census. The data is not cleaned. Although data is old, it can help to learn the regression techniques. The samples are given as 20641 rows and 10 columns of raw data. There are 10 columns of self-explanatory features that are shown in Table 1:

Table 1. Dataset: Features

Longitude	A measure of how far west a house is; a higher value is farther west
Latitude	A measure of how far north a house is; a higher value is farther north
housing_median_age	Median age of a house within a block; a lower number is a newer building
total_rooms	Total number of rooms within a block
total_bedrooms	Total number of bedrooms within a block
population	Total number of people residing within a block
households	Total number of households, a group of people residing within a home unit, for a block
median_income	Median income for households within a block of houses (measured in tens of thousands of US Dollars)
median_house_value	Median house value for households within a block (measured in US Dollars)
ocean_proximity	Location of the house w.r.t ocean/sea

All data is numerical, with the exception of ocean_ proximity, which has string input like ("near bay", "near" ocean", "inland"). Figure 1 illustrates the first 4 training examples of the dataset.

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity
0	-122.23	37.88	41.0	880.0	129.0	322.0	126.0	8.3252	452600.0	NEAR BAY
1	-122.22	37.86	21.0	7099.0	1106.0	2401.0	1138.0	8.3014	358500.0	NEAR BAY
2	-122.24	37.85	52.0	1467.0	190.0	496.0	177.0	7.2574	352100.0	NEAR BAY
3	-122.25	37.85	52.0	1274.0	235.0	558.0	219.0	5.6431	341300.0	NEAR BAY
4	-122.25	37.85	52.0	1627.0	280.0	565.0	259.0	3.8462	342200.0	NEAR BAY

Fig. 1. DataFrame: California Housing Prices

This problem is not supervised, because it doesn't have necessary labels for training. This is a regression type problem and the dataset is suitable for the regression modelling and quantity prediction.

The chosen dataset used for this unsupervised problem definitely has interesting variables that are correlated with pricing, some of them might have a greater effect while others less effect on prices of housing. For example, we can expect that for housing median age, the lower the number the higher the price and maybe the lower the population in the area the higher the price, as it can be considered a more private area. So it would be interesting to analyze these variables and their relations with prices, as well as create more accurate modeling and predicting of the houses prices based on the combination of these variables.

3 METHODS

In this project, we will be solving a machine learning problem. We will have an unsupervised problem, because the data does not come with labels. The technique that we will use to analyze the data will be linear regression. We will be using different algorithms to be able to do this. For example Random Forest Regression and Lasso. We will start by loading the data and dividing it into testing and training data. After the data is loaded, we will see how the categories/ features are correlated to the price using different tools. Some examples of possible tools are scatter plots, bar plots, etc. Then, we will pre-process the data, for example, this includes converting categorical values into numerical ones (categorical feature transformation), normalizing the data, etc. Then, we will be using the tools available with scikit-learn to fit the data. We will be using the fit method after creating an object using the classes from the different algorithms. Then, we will evaluate the model's performance using the predict method on the test set, present in the previously mentioned classes. Another possible way to evaluate the model is the Mean Squared Error. These tools will help us determine how successful our model is, and to adjust accordingly. We will optimize our model using hyperparameters tuning and experimenting with different feature combinations to adjust our model.

4 CONCLUSION

As most of us want to work in the industry sector once we finish our studies, this type of project equips us with valuable machine learning tools that we can use later on in our professional lives. For example, this project will apply regression models, which is a valuable tool recently learnt in this class. Predicting prices, as mentioned in the introduction, is extremely valuable. This project will help us see which methods are most effective and what information needs to be considered to make a proper prediction. Under-performing models can help us find which parameters are directly affecting the prices and which ones are not. Taking this into consideration will help us determine which combination of parameters will produce the best model. If this study is successful this model can be used for housing prices predictions in other regions. And the structure of the code can be followed to be applied in other projects related to pricing.

REFERENCES

[1] Bruno Afonso, Luckeciano Melo, Willian Oliveira, Samuel Sousa, and Lilian Berton. 2019. Housing Prices Prediction with a Deep Learning and Random Forest Ensemble. In *Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional* (Salvador). SBC, Porto Alegre, RS, Brasil, 389–400. https://doi.org/10.5753/eniac.2019.9300

[2] Okmyung Bin. 2004. A prediction comparison of housing sales prices by parametric versus semi-parametric regressions. *Journal of Housing Economics* 13, 1 (2004), 68–84. https://doi.org/10.1016/j.jhe.2004.01.001

[3] A. Géron. 2019. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems.* O'Reilly Media. https://books.google.com/books?id=HHetDwAAQBAJ

[4] Chaitali Majumder. [n. d.]. *House price prediction using machine learning.* Retrieved November 5, 2021 from https://nycdatascience.com/blog/student-works/machine-learning/house-price-prediction-using-machine-learning-2/

[5] Byeonghwa Park and Jae Kwon Bae. 2015. Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert Systems with Applications* 42, 6 (2015), 2928–2934. https://doi.org/10.1016/j.eswa.2014.11.040

[6] Subham Sarkar. September 6, 2019. *Predicting House prices using classical machine learning and Deep Learning Techniques.* Retrieved November 5, 2021 from https://medium.com/analytics-vidhya/predicting-house-prices-using-classical-machine-learning-and-deep-learning-techniques-ad4e55945e2d