**Name – Siddhesh Kadam**

**Net ID – vz3642**

**Literature Review**

**Dataset: California Housing Prices**

**Median house prices for California districts derived from the 1990 census**.

**Dataset Link:** [**https://www.kaggle.com/datasets/camnugent/california-housing-prices**](https://www.kaggle.com/datasets/camnugent/california-housing-prices)

**Original book: “Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow”.**

**Original book link:** [**https://www.google.com/books/edition/\_/X5ySEAAAQBAJ?hl=en&gbpv=0&kptab=overview**](https://www.google.com/books/edition/_/X5ySEAAAQBAJ?hl=en&gbpv=0&kptab=overview)

**Book Details:** SBN:9781098122461, 1098122461

Page count:864

Published: October 4, 2022

Publisher: [O'Reilly Media](https://www.google.com/search?hl=en&gbpv=1&pg=PT10&dq=Aur%C3%A9lien+G%C3%A9ron%27s+recent+book+%27Hands-On+Machine+learning+with+Scikit-Learn+and+TensorFlow%27&q=inpublisher:%22O%27Reilly+Media%22&tbm=bks&sa=X&ved=2ahUKEwj9uqGMnIuAAxXKJUQIHU_3BVUQmxMoAHoECCQQAg&sxsrf=AB5stBgrS7v6f1xrKHiZ4PVu1WZzYBrgqA:1689235359601)

Language: [English](https://www.google.com/search?hl=en&sxsrf=AB5stBgrS7v6f1xrKHiZ4PVu1WZzYBrgqA:1689235359601&q=&si=ACFMAn8gKcw6L8cgu575hwcueyKgoe9ZYQ1m1yJZTLKy7VGJ2kkUwOWyzurRofv6FPYFKkMDcrO562e9VU7keSOcLpOiXqDAunTlF6hf-yRsC606O2Z5xOxeWpoP_YjIQRs6HBDVB_IeLCHYarpHUrjjZPYjSh05l5Ib5TAagWArdmKoURZqtf8Rbz-6oQn34b0TD7T8R5-O&sa=X&ved=2ahUKEwj9uqGMnIuAAxXKJUQIHU_3BVUQmxMoAHoECCAQAg)

Author: [Aurélien Géron](https://www.google.com/search?hl=en&gbpv=1&pg=PT10&dq=Aur%C3%A9lien+G%C3%A9ron%27s+recent+book+%27Hands-On+Machine+learning+with+Scikit-Learn+and+TensorFlow%27&q=inauthor:%22Aur%C3%A9lien+G%C3%A9ron%22&tbm=bks&sa=X&ved=2ahUKEwj9uqGMnIuAAxXKJUQIHU_3BVUQmxMoAHoECCgQAg&sxsrf=AB5stBgrS7v6f1xrKHiZ4PVu1WZzYBrgqA:1689235359601)

**Book Summary: “**Hands-On Machine Learning with Scikit-Learn and TensorFlow" by Aurélien Géron is a bestselling book that introduces readers to deep learning and its impact on machine learning. The book emphasizes practical implementation using Python frameworks like Scikit-Learn, Keras, and TensorFlow. It provides concrete examples, minimal theory, and real-world Python code to help readers develop an intuitive understanding of building intelligent systems. The updated third edition covers techniques from simple linear regression to deep neural networks. The book also explores various models, unsupervised learning techniques, and neural network architectures such as convolutional nets, recurrent nets, generative adversarial networks, autoencoders, diffusion models, and transformers. It guides readers through building and training neural networks for applications in computer vision, natural language processing, generative models, and deep reinforcement learning. The book is designed to be accessible to programmers with minimal programming experience, making it a valuable resource for learning and applying machine learning techniques.

**Dataset Brief:**

This dataset is a modified version of the California Housing dataset available from [Luís Torgo's page](http://www.dcc.fc.up.pt/~ltorgo/Regression/cal_housing.html) (University of Porto). Luís Torgo obtained it from the StatLib repository (which is closed now). The dataset may also be downloaded from StatLib mirrors.

This dataset appeared in a 1997 paper titled Sparse Spatial Autoregressions by Pace, R. Kelley and Ronald Barry, published in the Statistics and Probability Letters journal. They built it using the 1990 California census data. It contains one row per census block group. A block group is the smallest geographical unit for which the U.S. Census Bureau publishes sample data (a block group typically has a population of 600 to 3,000 people).

**Dataset Changes:**

The dataset in this directory is almost identical to the original, with two differences:

* 207 values were randomly removed from the total bedroom’s column, so we can discuss what to do with missing data.
* An additional categorical attribute called ocean proximity was added, indicating (very roughly) whether each block group is near the ocean, near the Bay area, inland or on an island. This allows discussing what to do with categorical data.

**Project Objectives – *Python***

1. **Data splitting**: Splitting the data into training and testing sets to evaluate the performance of machine learning models accurately.
2. **Data preprocessing**: Cleaning the data by handling missing values in the "total\_bedrooms" column. Engineering new features, such as "income\_cat," by transforming the "median\_income" column and creating bins.
3. **Data visualization**: Using scatter plots and color maps to visualize the geographical distribution of housing data. Overlaying an image on scatter plots to provide geographical context. Correlation analysis: Examining the correlation between different features and the target variable, "median\_house\_value." Feature engineering: Creating new features, including "rooms\_per\_household," "bedrooms\_per\_room," and "population\_per\_household," based on existing features.
4. **Data description and summary**: Generating descriptive statistics of the dataset to gain insights into its distribution and variability.
5. **Data imputation**: Handling missing values in the dataset using techniques such as dropping rows with missing values, dropping the column, or filling missing values with the median.
6. **Data transformation**: Applying transformations to the dataset, including filling missing values with the median using the Simple Imputer and transforming numerical features using StandardScaler.
7. **Categorical data handling**: Extracting categorical features, such as "ocean\_proximity," for further analysis and encoding.

Encoding categorical features using techniques like OrdinalEncoder and OneHotEncoder.

1. **Custom transformer**: Creating a custom transformer, CombinedAttributesAdder, to add additional attributes to the dataset based on existing features.
2. **Data pipeline**: Creating a pipeline to streamline and automate the data transformation steps, including imputation, feature addition, and standardization.

**PROJECT SCOPE**

* Model Building and Evaluation: Implement various machine learning algorithms, such as regression models, decision trees, or ensemble methods, to predict housing prices. Evaluate their performance using appropriate metrics and compare the results.
* Feature Importance: Analyze the importance of different features in predicting housing prices. Identify the most influential factors that impact median house values in California.
* Advanced Visualization Techniques: Explore additional visualization techniques, such as interactive maps, heatmaps, or geospatial clustering, to gain deeper insights into the geographical patterns and relationships between variables.
* Outlier Detection and Handling: Investigate outliers in the dataset and analyze their impact on model performance. Implement outlier detection techniques and evaluate the effectiveness of outlier handling methods.
* Advanced Data Imputation: Explore advanced imputation techniques, such as regression-based imputation or iterative imputation, to handle missing values more accurately and assess their impact on model performance.
* Model Optimization: Fine-tune the machine learning models using hyperparameter optimization techniques to improve their predictive capabilities and achieve better performance.
* Advanced Feature Engineering: Experiment with different feature engineering techniques, such as polynomial features, interaction terms, or dimensionality reduction, to enhance the predictive power of the models.
* Ensemble Methods: Explore ensemble learning techniques, such as stacking, bagging, or boosting, to combine multiple models and improve the accuracy and robustness of predictions.

**Related Research Study Papers**

1. **A machine learning approach to big data regression analysis of real estate prices for inferential and predictive purposes**

**Link**: <https://www.tandfonline.com/doi/full/10.1080/09599916.2019.1587489?scroll=top&needAccess=true&role=tab>

**Summary:** This article proposes a machine learning approach to analyze big data, specifically real estate prices, for both inference and prediction purposes. While hedonic price regressions have typically been used for inference, machine learning applied to large datasets has the potential for accurate prediction. To address the integration of these strategies, the article introduces a novel methodology called 'incremental sample with resampling' (MINREM) for variable selection. The methodology is tested on two cases: web advertisements selling used homes in Colombia (61,826 observations) and data from the Metropolitan American Housing Survey 2011 (58,888 observations). The methodology consists of two stages: variable selection using MINREM and a traditional training and validation process for machine learning, with three additional activities. The results demonstrate the methodology's value in obtaining parsimonious and stable models across different sample sizes, while effectively combining inference and prediction. This paper presents an original methodology for regression analysis of big data.

**Roles in my project**:

* I will familiarize myself with the MINREM methodology described in the article. I aim to understand the variable selection process and the stages involved in developing regression models using machine learning techniques.
* Next, I will adapt the MINREM methodology to my specific research on California housing prices. I will modify the variable selection process and the training/validation procedures to suit my dataset and research objectives.
* With the adapted methodology, I will apply it to my dataset of California housing prices. I will implement the variable selection process using MINREM and develop regression models using machine learning techniques.
* Once I have the regression models, I will evaluate their performance for both inferential and predictive purposes. I will assess their stability, parsimony, and interpretability. I will also compare the results with other existing approaches or models.
* I will analyze the insights gained from the regression models and draw conclusions based on my research scopes. I will assess the accuracy of predictions and the importance of variables in determining housing prices in California.
* In my discussion, I will highlight the results and findings in the context of the article's contribution to big data regression analysis. I will emphasize any novel or unique aspects of my methodology compared to the one proposed in the article.

By implementing the MINREM methodology described in the article and adapting it to my research scopes, I aim to explore the potential of machine learning techniques in analyzing and predicting housing prices in California. This will allow me to contribute to the integration of hedonic price regressions and machine learning approaches and advance the field of big data regression analysis.

1. **Does machine learning prediction dampen the information asymmetry for non-local investors?**

**Link**: <https://journals.vilniustech.lt/index.php/IJSPM/article/view/17590>

**Summary:** This study investigates the predictive accuracy of machine learning methods, such as Random Forest, Gradient Boosting Machine, Support Vector Machine, and Deep Neural Networks, in estimating commercial real estate transaction prices. Using a dataset of 19,640 office properties from Costar spanning the 2004-2017 period in 10 major U.S. Consolidated Metropolitan Statistical Areas, the study compares the relative performance of these methods and identifies the most influential model for each office market. Partial dependence plots are employed to analyze the impact of research variables on predicted commercial office property values. The findings are expected to provide critical determinants for commercial office prices, surpassing traditional valuation models in terms of predictive power and offering valuable insights for out-of-state investors in understanding regional commercial real estate markets.

**Roles in my project**:

* Model Building and Evaluation: I will implement machine learning methods such as Random Forest, Gradient Boosting Machine, Support Vector Machine, and Deep Neural Networks to estimate housing prices in California. I will evaluate their predictive accuracy and compare their performance to identify the most effective model for predicting median house values.
* Feature Importance: I will utilize the methods presented in the article, such as partial dependence plots, to analyze the impact of different features on predicted housing prices in California. This will help me identify the most influential variables that contribute to the variations in median house values.
* Advanced Visualization Techniques: I will implement the visualization techniques discussed in the article, such as partial dependence plots and other suitable visualizations, to gain insights into the relationships between variables and geographical patterns of housing prices in California.
* Outlier Detection and Handling: I will adapt the outlier detection techniques used in the article to identify outliers in my housing price dataset. I will assess their impact on model performance and explore appropriate methods for handling outliers in the context of my project.
* Advanced Data Imputation: I will incorporate the advanced data imputation techniques described in the article, such as those used for handling missing values in the commercial real estate dataset, to improve the accuracy and completeness of my housing price dataset. I will evaluate the effectiveness of these techniques on the predictive performance of my models.
* Advanced Feature Engineering: I will experiment with advanced feature engineering techniques mentioned in the article, such as polynomial features or dimensionality reduction, to enhance the predictive power of my models for housing price estimation in California.
* By incorporating the findings and methodologies from the mentioned article, I aim to enhance the accuracy and predictive power of my models for estimating housing prices in California. I believe that these approaches will provide valuable insights into the determinants and dynamics of the regional real estate market, benefiting both local and out-of-state investors.

1. **Machine-Learning-Based Prediction of Land Prices in Seoul, South Korea**

**Link**: <https://www.mdpi.com/2071-1050/13/23/13088>

**Summary:** Accurately estimating real estate values is crucial for developing effective real estate policies that can navigate the complexities and volatility of the market. While statistical methods have been traditionally used for value estimation, machine learning methods have gained popularity due to their superior predictive capabilities. However, existing studies often overlook the separation of building and land prices when estimating real estate values using machine learning. In this study, we address this limitation by leveraging a vast amount of land-use information derived from diverse land and building datasets to estimate land prices. Specifically, we employ the random forest and XGBoost methods to estimate 52,900 land prices in Seoul, South Korea, from January 2017 to December 2020. We train separate models for different land uses and time periods. The results indicate that XGBoost exhibits higher prediction accuracy overall. Interestingly, when analyzing residential areas, the XGBoost models outperform the random forest models in terms of accuracy for the 2020 data, while the random forest models demonstrate better accuracy for the 2017-2020 data. Future analysis will expand the prediction model to incorporate submarkets defined by price volatility and locality, further enhancing its capabilities.

**Roles in my project**:

* Model Building and Evaluation: I can implement the random forest and XGBoost methods discussed in the article to build predictive models for estimating housing prices. I will train separate models for different land uses and time periods, similar to the study conducted in Seoul, South Korea.
* Feature Importance: Following the approach in the article, I can analyze the importance of different features, including land-use information, in predicting housing prices. This will help me identify the most influential factors that impact housing values in my specific context, such as the housing market in California.
* Advanced Visualization Techniques: Inspired by the article, I can explore advanced visualization techniques like interactive maps, heatmaps, or geospatial clustering to visualize the relationships between land-use variables and housing prices. This will provide me with deeper insights into geographical patterns and help me understand the impact of different factors on housing values.
* Outlier Detection and Handling: I can adapt the outlier detection techniques discussed in the article to identify and handle outliers in the housing price dataset. This will enable me to assess the impact of outliers on model performance and explore appropriate strategies for handling them effectively.
* Advanced Data Imputation: Building on the insights from the article, I can explore advanced data imputation techniques to handle missing values in the dataset. I can experiment with methods like regression-based imputation or iterative imputation to improve the accuracy of my models and evaluate their impact on predictive performance.
* Model Optimization: Inspired by the article's use of the XGBoost method, I can focus on optimizing the hyperparameters of my models using techniques like grid search or random search. This will allow me to fine-tune the models and improve their predictive capabilities for estimating housing prices.
* Advanced Feature Engineering: I can experiment with advanced feature engineering techniques, such as creating interaction terms or incorporating dimensionality reduction methods, to enhance the predictive power of my models. This will involve adapting the strategies discussed in the article to my specific housing price dataset.
* Ensemble Methods: Similar to the article's approach, I can explore ensemble learning techniques like stacking, bagging, or boosting to combine multiple models and improve the accuracy and robustness of my housing price predictions. This will involve implementing and evaluating different ensemble methods to find the most effective approach for my specific project.

By incorporating the insights and methodologies from the article into my project, I can enhance the accuracy of housing price estimation, gain valuable insights into the factors influencing housing values in California, and contribute to the development of effective real estate policies in the region.

1. **Effectiveness comparison of the residential property mass appraisal methodologies in the USA**

**Link**: <https://www.emerald.com/insight/content/doi/10.1108/17538271111153013/full/html>

**Summary:** This paper compares the prediction accuracy of three commonly used models, namely multiple regression, additive nonparametric regression, and artificial neural networks (ANN), for the mass appraisal of real estate. The study utilized a housing database of 33,342 residential houses, with a cutoff point of $88 per square foot for higher-priced homes. The findings suggest that both statistical models and ANN are reliable and cost-effective methods for mass appraisal of residential housing. Furthermore, while all three models show similar accuracy for lower and medium-priced houses, the ANN model demonstrates significantly higher accuracy for higher-priced houses. This research contributes to understanding the suitability and performance of different models in real estate mass appraisal.

**Roles in my project**:

* Implement multiple regression, additive nonparametric regression, and artificial neural networks (ANN) models for predicting housing prices.
* Compare the prediction accuracy of these models using appropriate evaluation metrics.
* Assess the accuracy of the models for different price ranges, particularly focusing on higher-priced houses.
* Analyze the cost-effectiveness and reliability of the statistical models and ANN for mass appraisal of residential housing.
* Incorporate advanced visualization techniques like interactive maps, heatmaps, or geospatial clustering to gain insights into geographical patterns and relationships between variables.
* Investigate outliers in the dataset and evaluate the impact on model performance. Implement outlier detection techniques and assess different outlier handling methods.
* Explore advanced data imputation techniques such as regression-based imputation or iterative imputation to handle missing values more accurately. Evaluate the impact of these techniques on model performance.
* Fine-tune machine learning models using hyperparameter optimization techniques to improve predictive capabilities. Experiment with different optimization algorithms and evaluate their impact on model performance.
* Experiment with advanced feature engineering techniques like polynomial features, interaction terms, or dimensionality reduction to enhance the predictive power of the models.
* Explore ensemble learning techniques such as stacking, bagging, or boosting to combine multiple models and improve the accuracy and robustness of predictions.

1. **Predicting Home Value in California, United States via Machine Learning Modeling**

Link: <http://iapress.org/index.php/soic/article/view/2019-M-5>

**Summary:** This paper addresses the challenge of predicting real estate market values by utilizing both linear and non-linear machine learning methods. Using real estate data from three counties in Los Angeles, California, the study aims to improve the accuracy of home value predictions. The findings indicate that traditional linear models are ineffective for complex datasets, while tree-based non-linear models offer the highest accuracy with the lowest mean square errors, highlighting their superiority for real estate value prediction.

**Roles in my project**:

* Incorporate tree-based non-linear models: Based on the article's findings, enhance your project by including tree-based non-linear models, such as decision trees or ensemble methods, to improve the accuracy of real estate value predictions compared to traditional linear models.
* Use appropriate evaluation metrics: Follow the article's recommendation of using suitable metrics, such as mean square error, to evaluate the performance of your models. This ensures a comprehensive assessment and enables effective comparison between different models.
* Analyze feature importance and visualize relationships: Apply feature importance analysis techniques to identify influential factors impacting housing prices in California. Additionally, leverage advanced visualization techniques like interactive maps, heatmaps, or geospatial clustering to gain deeper insights into geographical patterns and visualize variable relationships.

1. **Predicting property prices with machine learning algorithms**

Link: <https://www.tandfonline.com/doi/full/10.1080/09599916.2020.1832558>

**Summary:** This study evaluates the performance of three machine learning algorithms (SVM, RF, GBM) in property price appraisal using a dataset of 40,000 housing transactions in Hong Kong spanning over 18 years. RF and GBM outperform SVM in terms of predictive power, as indicated by metrics such as mean squared error (MSE), root mean squared error (RMSE), and mean absolute percentage error (MAPE). However, SVM is still useful for data fitting due to its ability to produce reasonably accurate predictions within a tight time constraint. Overall, machine learning shows promise as an alternative technique for property valuation and price prediction.

**Roles in my project**:

* Model Building and Evaluation: Implement regression models, decision trees, and ensemble methods to predict housing prices in California. Evaluate their performance using appropriate metrics to determine the most effective models for accurate predictions.
* Feature Importance: Analyze the significance of different features in predicting housing prices, identifying the most influential factors that impact median house values in California. Gain insights into the drivers of real estate prices to inform decision-making.
* Advanced Techniques: Utilize advanced visualization methods such as interactive maps, heatmaps, and geospatial clustering to explore geographical patterns and relationships between variables. Implement outlier detection techniques, advanced data imputation, and model optimization to improve prediction accuracy and reliability.

1. **Predicting owner-occupied housing values using machine learning: an empirical investigation of California census tracts data**

Link: <https://www.tandfonline.com/doi/abs/10.1080/09599916.2021.1890187>

**Summary:** This paper introduces machine learning methods, including Ridge, LASSO, and Elastic Net regressions, to predict housing prices in the presence of multiple covariates. The empirical results demonstrate that these supervised ML methods outperform conventional OLS-based approaches in providing comprehensive descriptions of owner-occupied housing values in California's census tracts. The ML methods offer advantages such as variable selection, overcoming overfitting issues, and delivering improved out-of-sample predictions.

**Roles in my project**:

* Model Building and Evaluation: Implement a range of machine learning algorithms, including regression models, decision trees, ensemble methods, Ridge, LASSO, and Elastic Net regressions, to predict housing prices. Evaluate their performance using appropriate metrics and compare them to conventional OLS-based methods to identify the most effective models.
* Feature Importance: Analyze the determinants of owner-occupied housing values in California's census tracts by assessing the significance of different features. Identify the most influential factors impacting median house values to gain insights into the drivers of real estate prices and inform decision-making.
* Advanced Visualization Techniques: Utilize advanced visualization methods, such as interactive maps, heatmaps, and geospatial clustering, to explore geographical patterns and visualize the relationships between variables. Enhance understanding and interpretation of the data, enabling better insights into the factors influencing housing prices in California.