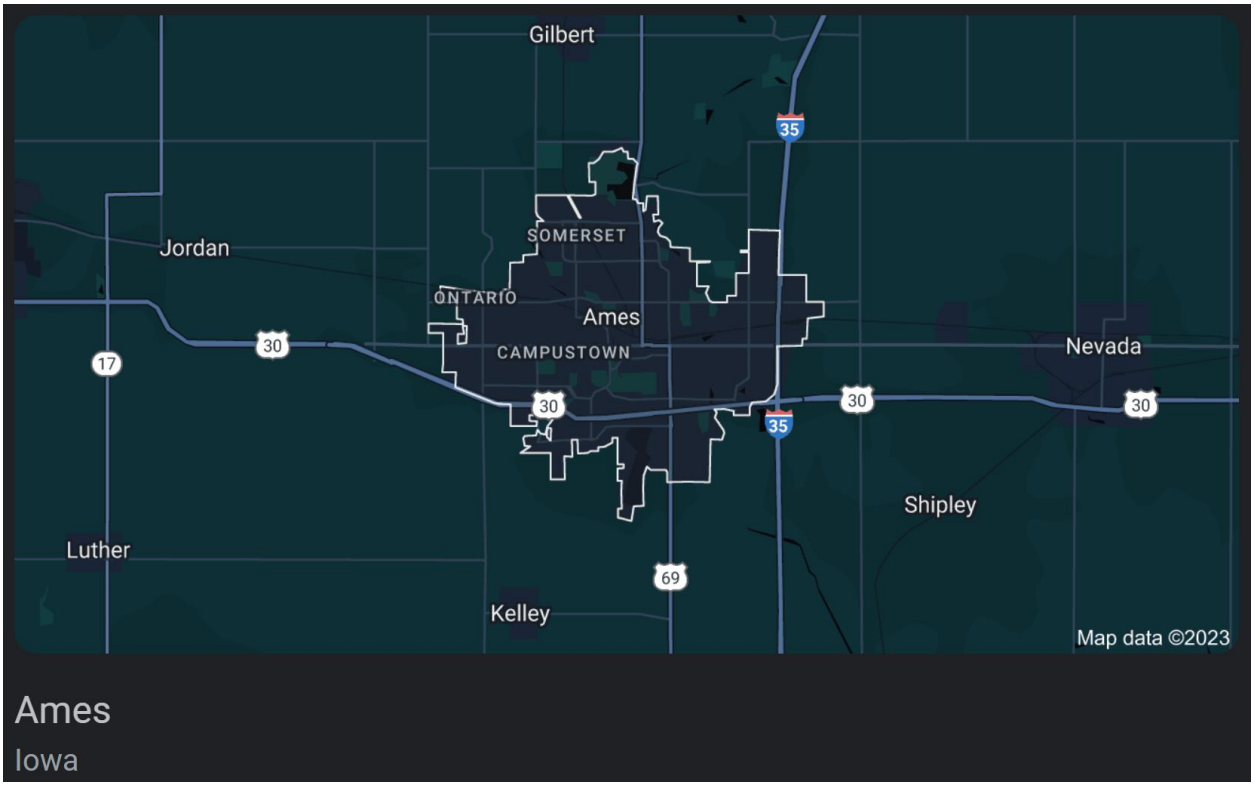
Predict the House Prices for Residential homes in Ames, Iowa



Springboard - DSC

Capstone Project 2

By Asiya Shakeel

July 2023

1. Introduction:

The House Price Prediction Capstone project aimed to develop a data-driven solution for accurately predicting house prices. The project targeted real estate agents, buyers, sellers, and investors as the intended stakeholders. By utilizing various data science techniques, including exploratory data analysis, feature engineering, and machine learning algorithms such as linear regression, random forest, and gradient boosting, the project successfully built predictive models. Additionally, the SHAP framework was employed to interpret the models and analyze the importance of different features in determining house prices. The project's detailed implementation, analysis, and deliverables can be found in the GitHub repository https://github.com/asiyashakeel78/Spring-Board/tree/main/Capstone%20Project%202

, providing stakeholders with valuable insights and informed decision-making tools in the real estate market.

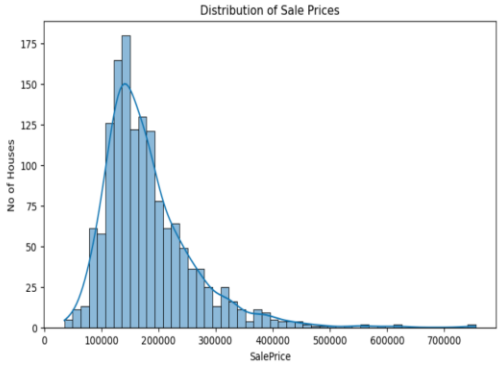
For the House Price Prediction Capstone project, the data used was obtained from Kaggle.com, a popular online platform for hosting data science competitions and datasets. The dataset was carefully selected to include various features related to residential properties, such as square footage, number of rooms, location, and amenities, which are crucial factors in determining house prices. The data acquisition process involved retrieving the dataset from Kaggle and ensuring its quality, integrity, and suitability for analysis. The next step was data wrangling, which involved handling missing values, outliers, and transforming variables as necessary to prepare the dataset for analysis. By performing these data acquisition and wrangling steps, we ensured that the dataset was in a suitable format for further exploration, modeling, and prediction of house prices in the subsequent stages of the project.

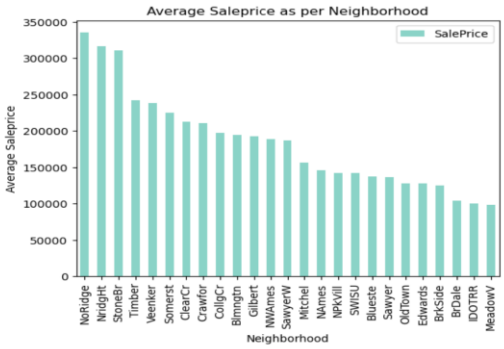
2. Approach:

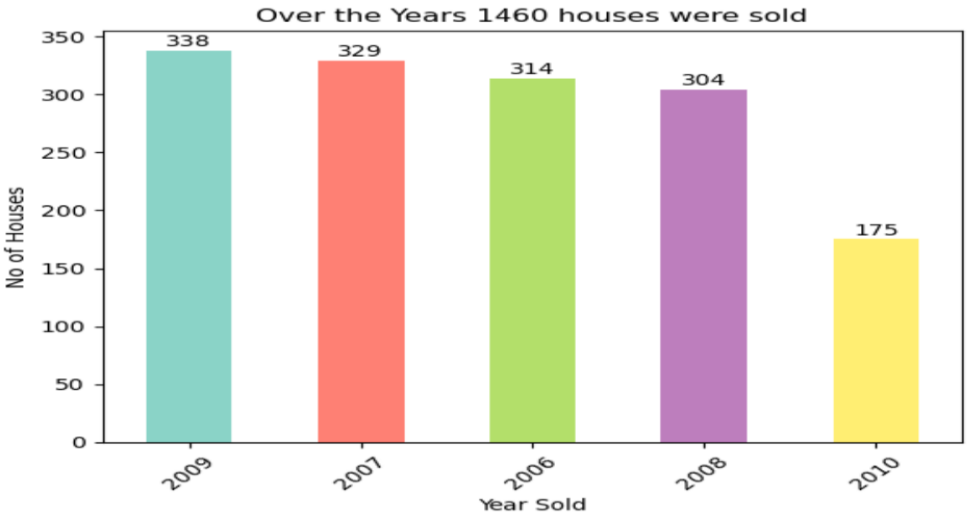
2.1 Data Acquisition and Wrangling:

For this project , the data used was obtained from Kaggle.com, a popular online platform for hosting data science competitions and datasets. The dataset was carefully selected to include various features related to residential properties, such as square footage, number of rooms, location, and amenities, which are crucial factors in determining house prices. The data acquisition process involved retrieving the dataset from Kaggle and ensuring its quality, integrity, and suitability for analysis. The next step was data wrangling, which involved handling missing values, outliers, and transforming variables as necessary to prepare the dataset for analysis. By performing these data acquisition and wrangling steps, we ensured that the dataset was in a suitable format for further exploration, modeling, and prediction of house prices in the subsequent stages of the project. The dataset contained information on 1460 houses located in 25 different neighborhoods.

2.2 Story Telling and Inferential Statistics:







The first plot reveals that the price range of houses in the dataset spans from $50,000 to $800,000. This demonstrates the wide variability in housing prices within the dataset. Moving on to the second plot, it highlights that the "NoRidge" neighborhood has the highest average house price among the 25 neighborhoods considered. Conversely, the "MeadowV" neighborhood exhibits the lowest average house price. These findings provide valuable insights into the variation in prices across different neighborhoods. Lastly, the third plot indicates that, on average, over 300 houses per year have been sold in the past years. This information suggests a relatively active housing market with a significant number of transactions taking place. By actively analyzing these plots, we gain a deeper understanding of the price range, neighborhood disparities, and market activity within the dataset.

2.3 Baseline Modeling:

During the baseline modeling phase, we implemented Linear Regression as the initial model, which resulted in a significantly high Mean Absolute Percentage Error (MAPE) of 3632.40. Our analysis revealed substantial discrepancies between the predicted and actual values, indicating the presence of non-linearity within the data. Additionally, the histograms depicting the distribution of residuals in the test set highlighted a wide spread of differences. In an effort to improve our results, we decided to explore the K-nearest neighbors (KNN) algorithm. Although the KNN algorithm yielded a lower MAPE of 0.179, demonstrating an improvement over the baseline model, we observed that there is still room for enhancing the accuracy of our predictions. The scatter plot and histograms emphasized the need for further refinement. As a next step, we plan to explore alternative algorithms, such as RandomForestRegressor, to bolster the overall performance of our house price prediction model.

2.4 Extended Modeling:

In the advanced modeling phase of our notebook, we aimed to enhance the accuracy of house price predictions by implementing three regression models: RandomForestRegressor, LGBMRegressor, and XGBRegressor. To optimize the performance of these models, we utilized two hyperparameter tuning tools, GridSearchCV and RandomizedSearchCV. Through a comprehensive analysis, we evaluated the models based on metrics such as Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE). The results indicated that the XGBRegressor model with GridSearchCV tuning consistently outperformed the other models, achieving the lowest MAE of $16,718 and MAPE of 11%. This finding highlights the XGBRegressor model with GridSearchCV tuning as the most accurate and reliable approach for predicting house prices.

3. Findings:

Model	MAE	MAPE
RandomForestRegressor	$19,679	12%
LGBMRegressor	$17,976	11%
XGBRegressor	**$18,707**	**12%**
RandomForestRegressor (GridSearch)	$19,455	12%
LGBMRegressor (GridSearch)	$16,758	11%
XGBRegressor (GridSearch)	**$16,718**	**11%**
RandomForestRegressor (RandomSearch)	$19,814	12%
LGBMRegressor (RandomSearch)	$18,152	11%
XGBRegressor (RandomSearch)	$17,261	11%

After evaluating the models and tuning them using GridSearchCV and RandomizedSearchCV, we have chosen XGBRegressor with GridSearch tuning as our primary model. It achieved the lowest MAE of $16,718 and MAPE of 11%. This model provides the most accurate and reliable predictions for house prices based on the given test dataset.

4. Conclusions and Future Work:

In conclusion, the House Price Prediction Capstone project successfully developed a predictive model for house prices using advanced modeling techniques. By comparing and evaluating various regression models, such as RandomForestRegressor, LGBMRegressor, and XGBRegressor, we identified the XGBRegressor model with GridSearch tuning as the most accurate and reliable for predicting house prices, achieving a low MAE of $16,718 and MAPE of 11%.

However, there is still room for improvement in enhancing the model's performance.

- Incorporating external datasets such as economic indicators, crime rates, or school ratings to capture additional factors that influence house prices.

- Future work could involve exploring additional feature engineering techniques, incorporating more advanced algorithms, or incorporating external data sources to further refine the predictive accuracy.

- Additionally, conducting a thorough analysis of influential outliers, addressing multicollinearity issues, and considering time series aspects could contribute to the model's overall effectiveness. Continued research and experimentation will pave the way for enhanced house price prediction models and improved decision-making in the real estate industry.

5. Recommendations for the clients:

- **Investment Opportunities:** Identify lucrative investment opportunities in the real estate market. The model's predictions can guide clients in identifying undervalued properties or areas with potential for growth, helping them make informed decisions and maximize their return on investment.
- **Risk Assessment:** Assess the risk associated with property investments. The model's accuracy in predicting house prices can aid in evaluating the potential risks and returns of investment projects, assisting clients in making informed decisions and mitigating financial risks.

6. Consulted Resources:

https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/discussion/163335

https://scikit-learn.org/stable/auto_examples/model_selection/plot_cv_predict.html#sphx-glr-auto-examples-model-selection-plot-cv-predict-py

https://seaborn.pydata.org/generated/seaborn.histplot.html

https://scikit-learn.org/stable/auto_examples/ensemble/plot_ensemble_oob.html#sphx-glr-auto-examples-ensemble-plot-ensemble-oob-p

https://scikit-learn.org/stable/user_guide.html

https://github.com/slundberg/shap

https://medium.com/dataman-in-ai/explain-your-model-with-the-shap-values-bc36aac4de3d

https://shap.readthedocs.io/en/latest/example_notebooks/tabular_examples/model_agnostic/Diabetes%20regression.html

https://towardsdatascience.com/explainable-ai-xai-with-shap-regression-problem-b2d63fdca670