Title: Predictive Insights into Property Valuation: Analyzing the Ames Housing Dataset

HEMANTH KURRA

**Introduction-**

The Ames Housing Dataset is a frequently utilized compilation of data within the realms of machine learning and data analysis. It provides an extensive array of information concerning residential properties in Ames, Iowa, serving as a valuable resource for analyzing housing market patterns and forecasting property values. With its inclusion of a wide range of numerical and categorical features, the dataset offers a detailed depiction of each property, facilitating in-depth exploration and analysis.

The Ames Housing Dataset originates from the city of Ames, Iowa, United States. It was compiled by Dean De Cock for the purpose of facilitating research and analysis in the fields of real estate, data science, and related areas. The dataset is often used as a benchmark in predictive modeling and machine learning tasks due to its comprehensive coverage of various attributes associated with residential properties in the Ames area.

The Ames Housing Dataset is indeed available on **Kaggle**, a popular platform for data science competitions and datasets. Kaggle hosts a variety of datasets contributed by users, and the Ames Housing Dataset is one of them. It was originally sourced from the city of Ames, Iowa, but it's made available and widely used through the Kaggle platform for analytical purposes, including predictive modeling, data visualization, and exploratory analysis.

**Project Proposal: Real Estate Price Prediction Application**

I proposed the development of an R Shiny-based Real Estate Price Prediction Application, tailored to assist homeowners in estimating the market value of their property with precision. This innovative tool will analyze user-provided property details—such as the number of bedrooms and bathrooms, garage availability, and construction year—against a comprehensive dataset of similar properties to forecast a competitive sale price. The application will offer an interactive platform displaying comparable recent sales, enabling users to perform a market analysis and understand pricing trends. Our aim is to deliver a data-driven, user-friendly application that simplifies the property valuation process, providing homeowners with clear, actionable insights for informed decision-making in the real estate market.

**Data Cleaning**-

We imported our dataset into RStudio and conducted various data cleaning tasks. Initially, we wrote code to identify missing or null values in the dataset. Subsequently, we eliminated columns with high percentages of missing values, namely "Alley," "Pool.QC," "Fence," and "Misc.Feature," which collectively contained approximately 2600 missing values out of 2900 rows.

Regarding the "Lot Frontage" column, missing values were replaced with 0, implying properties with no linear feet of street connection. This approach, known as imputation, assigns a specific value to missing data based on assumptions or existing data patterns.

We then addressed null values in basement-related columns by replacing them with "No," ensuring consistency across these columns and indicating the absence of a basement or relevant feature.

For the "Mas.Vnr.Area" column, missing values were replaced with the mean value, calculated excluding the missing values. This imputation technique maintains dataset characteristics by substituting missing values with a representative statistic.

Further, we filtered rows where specific garage-related attributes ('Garage.Type,' 'Garage.Yr.Blt,' 'Garage.Finish,' 'Garage.Cond,' and 'Garage.Qual') had no missing values, facilitating analysis or modeling tasks.

The "FirePlace.Qu" column's missing values were replaced with "No," ensuring uniformity and indicating the absence of a fireplace or the corresponding quality.

After completing data cleaning, we created two new variables: "Total.Bathrooms," which sums the count of full bathrooms, half bathrooms, and their basement equivalents; and "Garage," indicating whether a property has a garage based on the presence of at least one parked car ("Garage.Cars" > 0).

**Key facts and Trends observed**-

After analyzing the Ames Housing Dataset, several significant observations and patterns emerged, providing insights into various aspects of residential properties in Ames, Iowa. Firstly, an apparent trend is evident in the distribution of property prices ('SalePrice'). The dataset displays a wide spectrum of prices, with some properties priced notably higher than others. This indicates the presence of diverse housing options catering to various budget preferences within the Ames real estate market. Additionally, although a majority of properties fall within a particular price range, there are outliers on both ends of the scale, suggesting the existence of luxury properties as well as more economical alternatives.

Another notable finding concerns the correlation between property age ('Year.Built') and sale price ('SalePrice'). Analysis reveals that newer properties generally command higher prices compared to older ones. This emphasizes the importance of property age as a determinant of pricing, with buyers often willing to pay more for newer constructions featuring contemporary amenities and designs. Furthermore, the dataset demonstrates a broad range of construction years, reflecting the historical evolution of housing stock in Ames over time. Understanding this temporal dimension is essential for assessing market dynamics and identifying investment or redevelopment opportunities.

Moreover, an intriguing pattern emerges regarding the presence of garages ('Garage') and its influence on property prices. Properties with garages tend to fetch higher prices than those without, indicating the value buyers attribute to this amenity. This underscores the significance of factors beyond the primary dwelling, such as additional features like garage space, in impacting property valuations. Additionally, the dataset offers insights into the prevalence of garages across different neighborhoods and property types, highlighting disparities in housing preferences and market demand. Overall, these key observations and trends provide valuable insights into the dynamics of the Ames housing market, guiding strategic decision-making for buyers, sellers, and investors alike.

**Predictive Models**-

We experimented with several predictive models during our analysis. Among the models considered were linear regression, decision trees, support vector machines, gradient boosting machines, and random forests. Each model was evaluated based on its performance metrics such as mean squared error, root mean squared error, and R-squared value.

Ultimately, we chose the Random Forest model as our final predictive model. This decision was based on several factors. Firstly, Random Forest is known for its robustness to overfitting and its ability to handle large datasets with high dimensionality. Additionally, it often performs well with minimal tuning of parameters. Moreover, the model's ability to capture nonlinear relationships and interactions between variables made it well-suited for our dataset, which included a mix of numerical and categorical features. Finally, the Random Forest algorithm provides a measure of variable importance, aiding in the interpretation of the model's predictions and identifying key features influencing the target variable, SalePrice, in our case. Overall, the Random Forest model emerged as the most suitable choice based on its performance and versatility for our dataset.

**Application and User Guide-**

We've developed an application using RShiny. The following screenshot serves as a user interface for obtaining property sale price estimates. This tool functions as a real estate price prediction tool, enabling users to input various parameters such as the number of bedrooms, bathrooms, and other relevant details to receive an estimated sale price for a property.

The application features a form where users can input information including the number of bedrooms, bathrooms, garage availability, year built, and more, subsequently predicting the sale price. Additionally, there's a table displaying recent sales data for users to compare against their property, aiding in understanding market value. Users can search for specific properties within this data table and filter results based on different criteria, allowing them to focus on properties similar to their own. Furthermore, the data table is paginated for easy navigation through numerous entries.A screenshot of a computer

Description automatically generated

Our application provides users with the flexibility to adjust their requirements based on the displayed price. It relies on user input regarding the desired number of bathrooms, bedrooms, garage inclusion, and the year of construction of the house. Upon submission, the application generates results corresponding to the user's inputs, showcasing the number of outputs produced for their housing needs.

Overall, This is a well-designed and informative RShiny application that could be very useful for people who are looking to buy or sell a home. It is easy to use, provides valuable data and insights, and has a clean and user-friendly interface.

**References-**

AMES Housing Dataset. (2023, June 20). Kaggle. <https://www.kaggle.com/datasets/shashanknecrothapa/ames-housing-dataset>

Organized processes to clean data. (n.d.). Data Science Stack Exchange. <https://datascience.stackexchange.com/questions/52/organized-processes-to-clean-data>

Shiny. (n.d.). Shiny. <https://shiny.posit.co/r/articles/start/build/>

M, S. (2023, February 23). Introduction to Random Forest in R. Simplilearn.com. <https://www.simplilearn.com/tutorials/data-science-tutorial/random-forest-in-r>

Maklin, C. (2022, May 9). Random Forest in R - towards data science. Medium. <https://towardsdatascience.com/random-forest-in-r-f66adf80ec9>