Alessandro Gatti

Final Project Report: Housing Price Prediction Model

1. Introduction

The goal of this project was to develop a predictive model that could accurately estimate house prices based on a range of features of properties in Ames, Iowa from the dataset. This task is crucial for real estate investors, home buyers, and analysts who need to evaluate property values in a dynamic market. Several machine learning techniques were explored, including linear regression, regularization techniques like Ridge Regression, and more complex models like Random Forest and XGBoost. The final output was a blended model of Ridge and Random Forest, aimed at balancing bias and variance to achieve more robust predictions.

This report outlines the process of the project, from data preprocessing to model evaluation, highlighting where visualizations were used to support insights. Each phase of the project was documented with visualizations to guide decision-making, which were essential to understanding feature relationships, detecting outliers, and selecting models.

2. Data Preprocessing

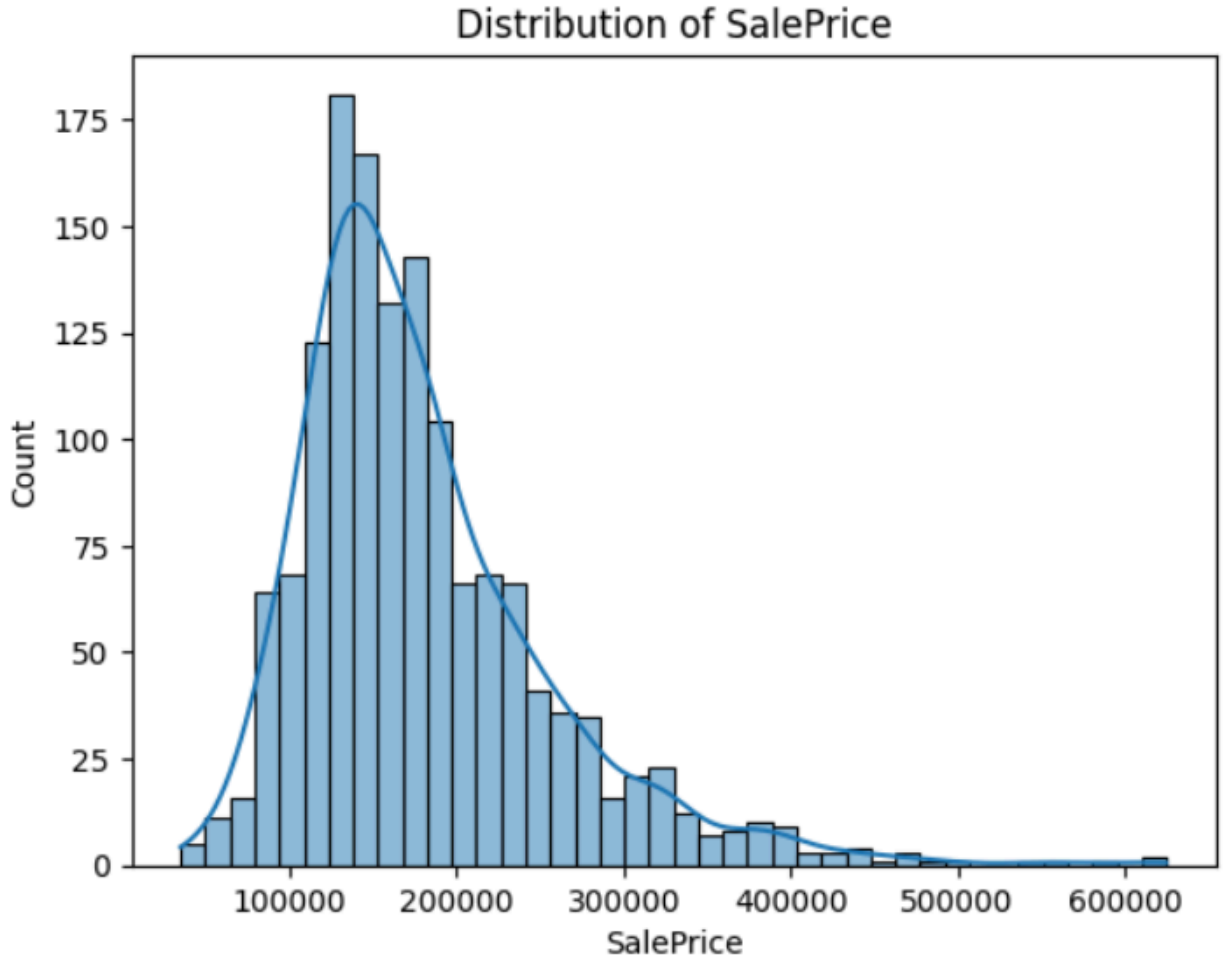
The dataset contained various features, including categorical and numerical variables, which required thorough preprocessing before feeding into machine learning models. The preprocessing involved several key steps:

- **Handling Missing Data:** Missing values were addressed appropriately. For categorical columns, missing data was handled by filling with the most frequent category or a specific 'missing' placeholder. For numerical columns, simple imputation methods were used.

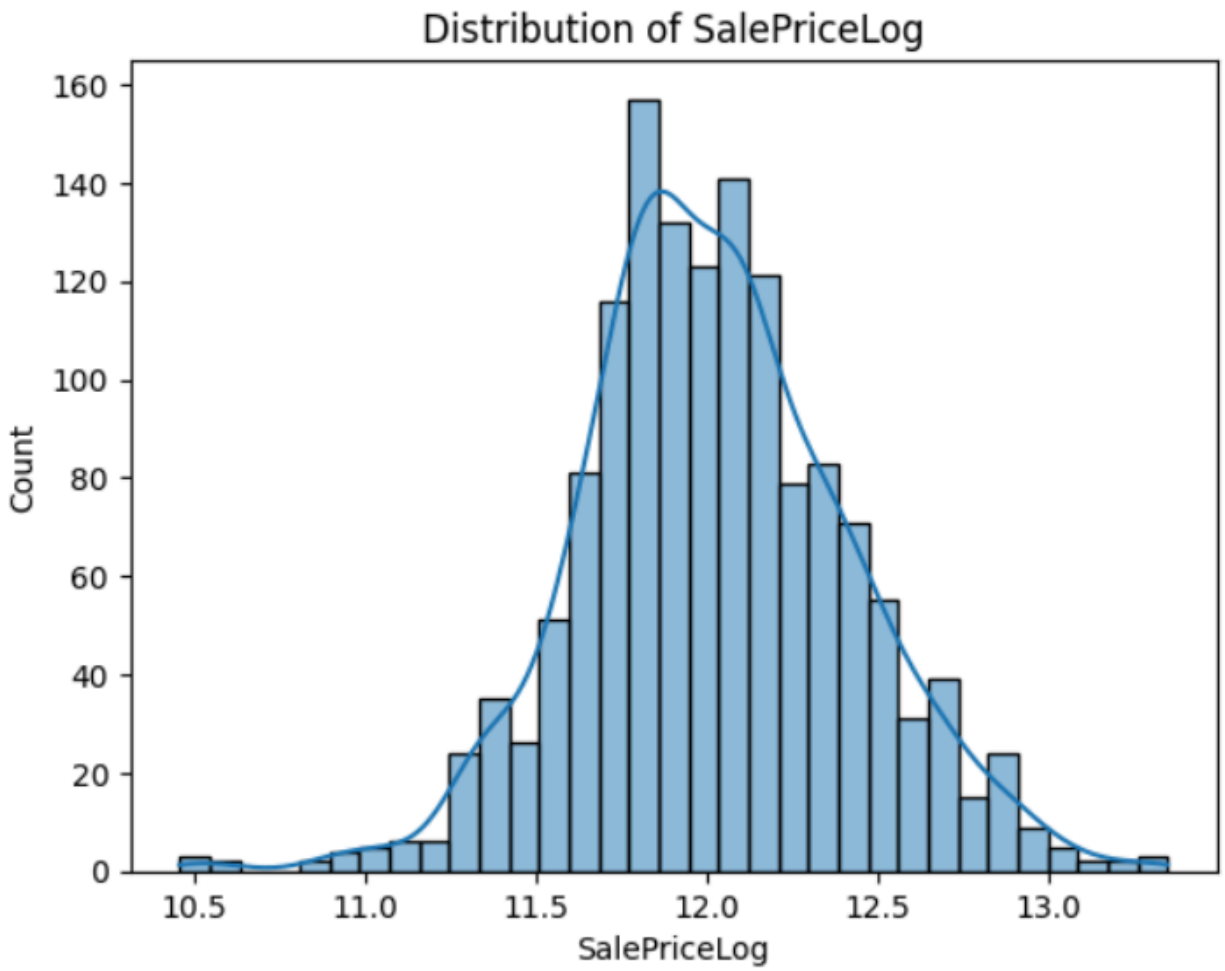
- **Encoding Categorical Variables:** Categorical features were transformed into numerical values using ordinal encoding for ordered categories and one-hot encoding for nominal categories. This allowed me to maintain the structure of the categorical variables while making them compatible with machine learning algorithms.

- **Feature Scaling and Normalization:** StandardScaler was used to normalize the numerical features. This step ensured that all features contributed equally to the model training, especially for models like Ridge that are sensitive to the scale of the input data.

- **Outlier Removal:** Outliers in key variables like 'GrLivArea' and 'TotalBsmtSF' were removed based on domain knowledge and visual analysis of histograms. This helped ensure the model was not skewed by extreme values. The removal of outliers was a necessary step for improving model stability.



The SalePrice distribution plot provides valuable insight into the distribution of house prices within the dataset. The histogram displays the frequency of various house prices, while the overlaid kernel density estimate (KDE) curve provides a smoothed representation of this distribution. From the visualization, it's clear that the distribution is right-skewed, with most house prices concentrated between $100,000 and $300,000. There are also fewer properties sold at higher price points, with a long tail extending toward $600,000. This visualization is crucial in understanding the general trend of house prices and helps inform our decision to apply a log transformation later to normalize the target variable for regression modeling.



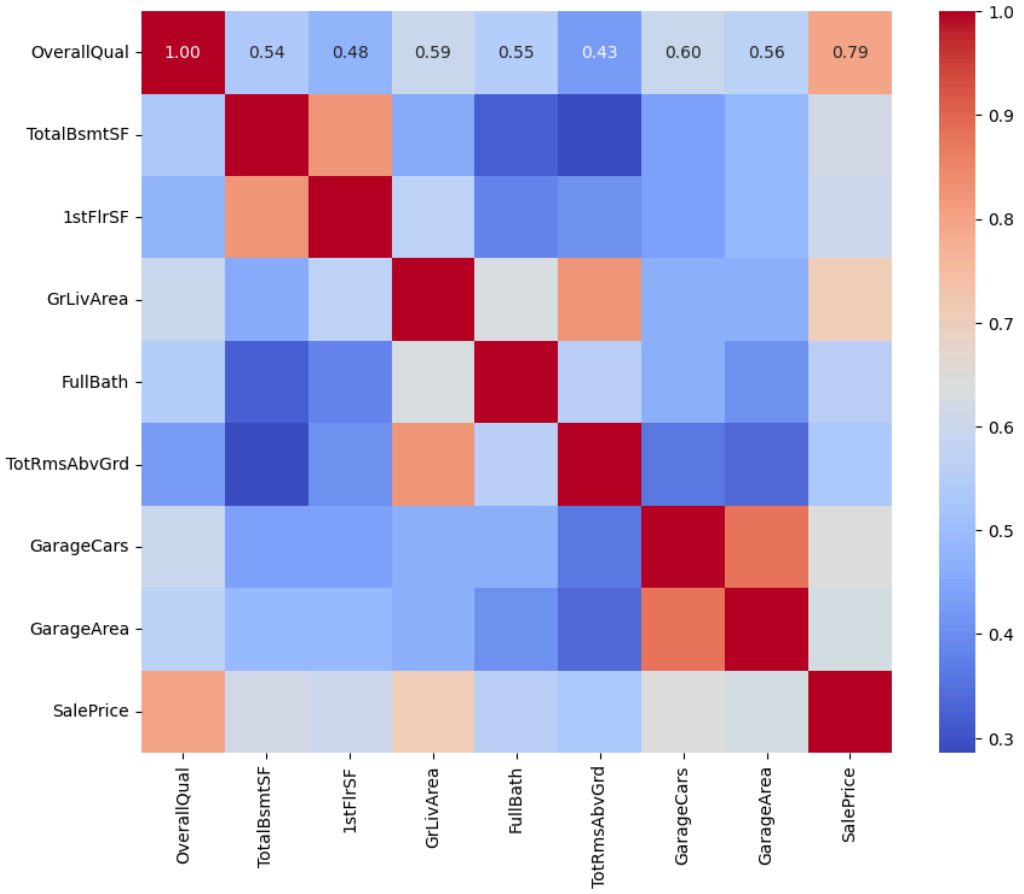
The Distribution of SalePriceLog visualization provides a detailed look at the logarithmic transformation of the sale prices of properties in our dataset. By applying the logarithmic transformation, the skewness observed in the original SalePrice distribution (as shown in the earlier graph) is corrected, bringing the data closer to a normal distribution. This transformation was necessary to meet the assumptions of linear regression models, which typically perform better when the target variable follows a normal distribution. The plot shows a more symmetric bell curve, with the peak around a SalePriceLog value of approximately 12, indicating that most property prices fall within a certain range after the log transformation.

3. Exploratory Data Analysis (EDA)

Exploratory Data Analysis was conducted to understand feature relationships and distributions. Correlation matrices and scatter plots were used to assess relationships between features like 'TotalSF', 'OverallQual', and 'SalePrice'.

- **Correlation Matrix:** This allowed me to identify highly correlated features. Features like 'GarageCars', 'OverallQual', and 'GrLivArea' showed strong positive correlations with

'SalePrice'. These features were given more weight in model building due to their significance.



This correlation heatmap shows the relationship between the *SalePrice* variable and the top numerical features in the dataset, such as 'OverallQual', 'TotalBsmtSF', '1stFlrSF', 'GarageCars', and others. The darker the red color, the stronger the positive correlation between the two variables, while the darker the blue color represents a stronger negative correlation. From the heatmap, 'OverallQual' has the highest correlation with 'SalePrice' (0.79), followed by features like 'GrLivArea', 'GarageCars', and 'TotalBsmtSF'. This visualization provides valuable insights into which features are most influential in predicting the sale price of homes. It helps inform model development, as I can focus on these high-correlation features for better predictions.

- **Feature Importance:** Features with a correlation lower than 0.2 with 'SalePrice' were considered for removal. This threshold helped streamline the dataset, ensuring that only relevant features were kept.

4. Model Development and Training

Several machine learning models were tested, including linear models like Ridge Regression and more complex models like Random Forest and XGBoost. The following models were developed and evaluated:

- **Ridge Regression:** A linear model with L2 regularization. Ridge was chosen due to its ability to handle multicollinearity between features while minimizing overfitting.

- **Random Forest:** This ensemble method leverages multiple decision trees to create a more robust model. It was chosen due to its flexibility and ability to capture non-linear relationships between features.

- **Blended Model:** Given the strengths and weaknesses of both Ridge Regression (linear relationships) and Random Forest (non-linear relationships), a blended model was created to combine predictions from both models, leveraging their strengths. This reduced the overall variance without increasing bias too much, resulting in more accurate predictions.

Model R-Squared and RMSE Values		
Model	RMSE	R-Squared
Ridge	20452.6866	0.9210
Random Forest	22600.4205	0.9035
Blended	20446.5387	0.9210

The table presented shows the RMSE (Root Mean Squared Error) and R-Squared values for the three models used: Ridge Regression, Random Forest, and a Blended Model. These metrics are essential for evaluating the model performance. RMSE provides a measure of the average difference between the predicted and actual values of SalePrice, with lower values indicating better fit. R-Squared indicates the proportion of variance in the dependent variable that is predictable from the independent variables, with higher values representing better predictive power.

In this case, the Ridge model and the Blended model yielded the best RMSE results, both around 20,446-20,452, with the Blended model slightly outperforming Ridge in terms of predictive accuracy by just a small margin. Both models achieved an R-Squared of 0.9210, indicating they explained 92.1% of the variance in SalePrice, which is a very strong result. The Random Forest model had a slightly higher RMSE of 22,600 and a lower R-Squared of 0.9035, showing that it was less accurate than the Ridge and Blended models. This suggests that blending models can provide a more robust solution by balancing the strengths of both Ridge and Random Forest, reducing overfitting while maintaining high predictive power.

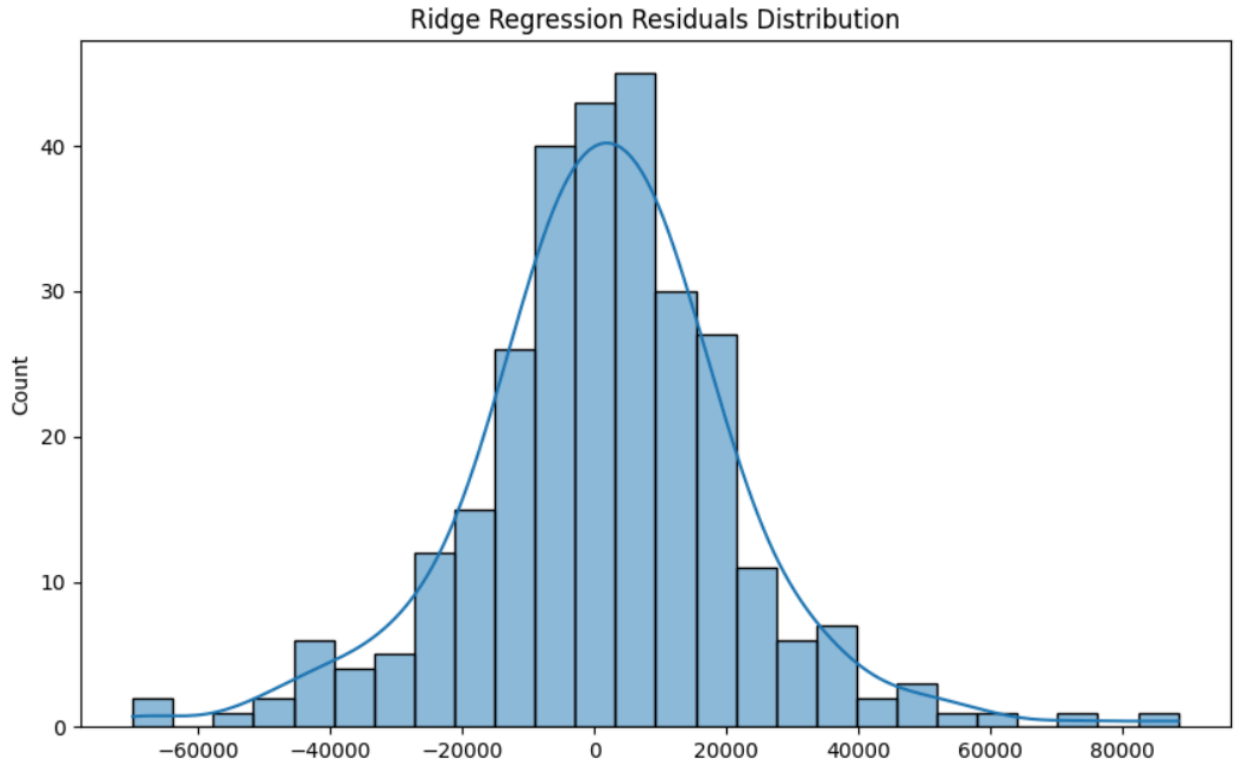
5. Model Evaluation

The models were evaluated using Root Mean Squared Error (RMSE) to measure the accuracy of the predictions. The models were compared using RMSE from cross-validation and test datasets.

- **Ridge Regression Results:** The RMSE on the test set was relatively low, indicating that the linear model was performing well on the dataset. However, the Ridge model struggled with non-linear relationships in the data.

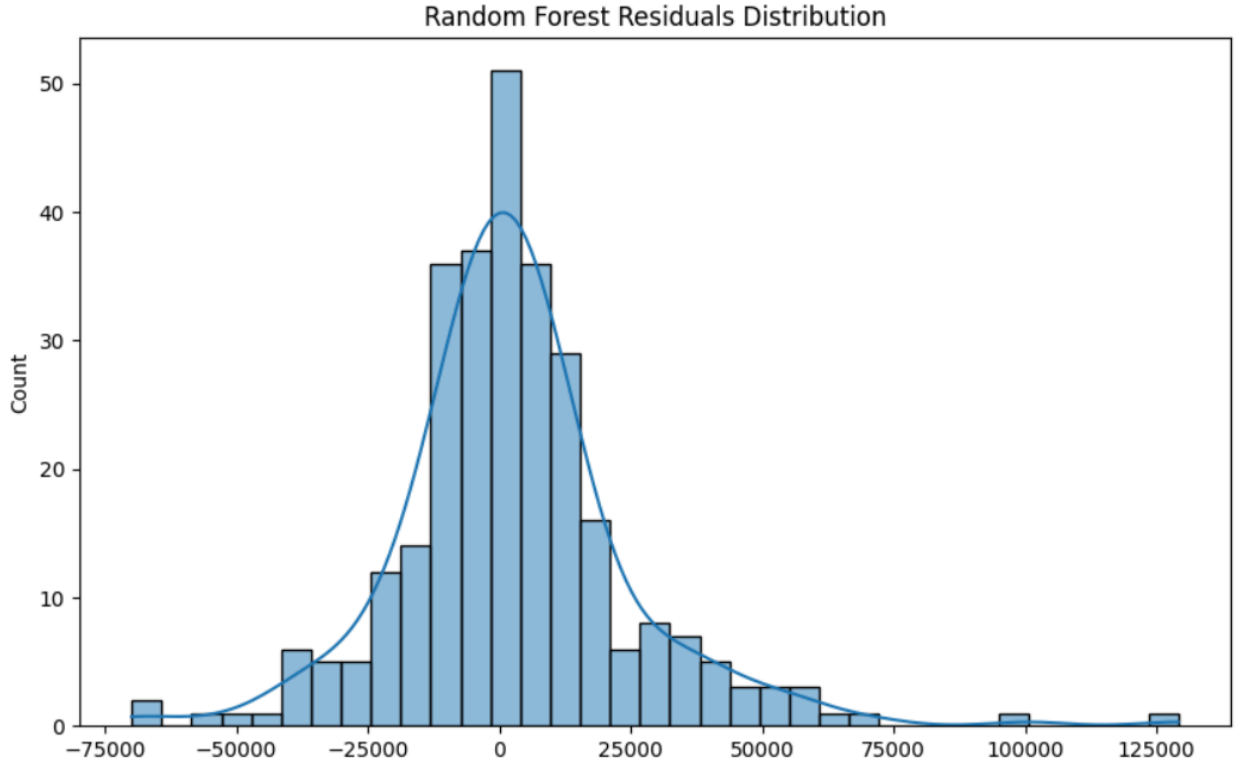
- **Random Forest Results:** The Random Forest model performed slightly better in capturing the complex relationships in the data but was prone to overfitting due to its complexity.

- **Blended Model:** The final blended model achieved the best RMSE score, striking a balance between bias and variance. By combining both Ridge and Random Forest predictions (57% Ridge, 43% RF), the model mitigated overfitting issues and delivered stable results across different subsets of data.



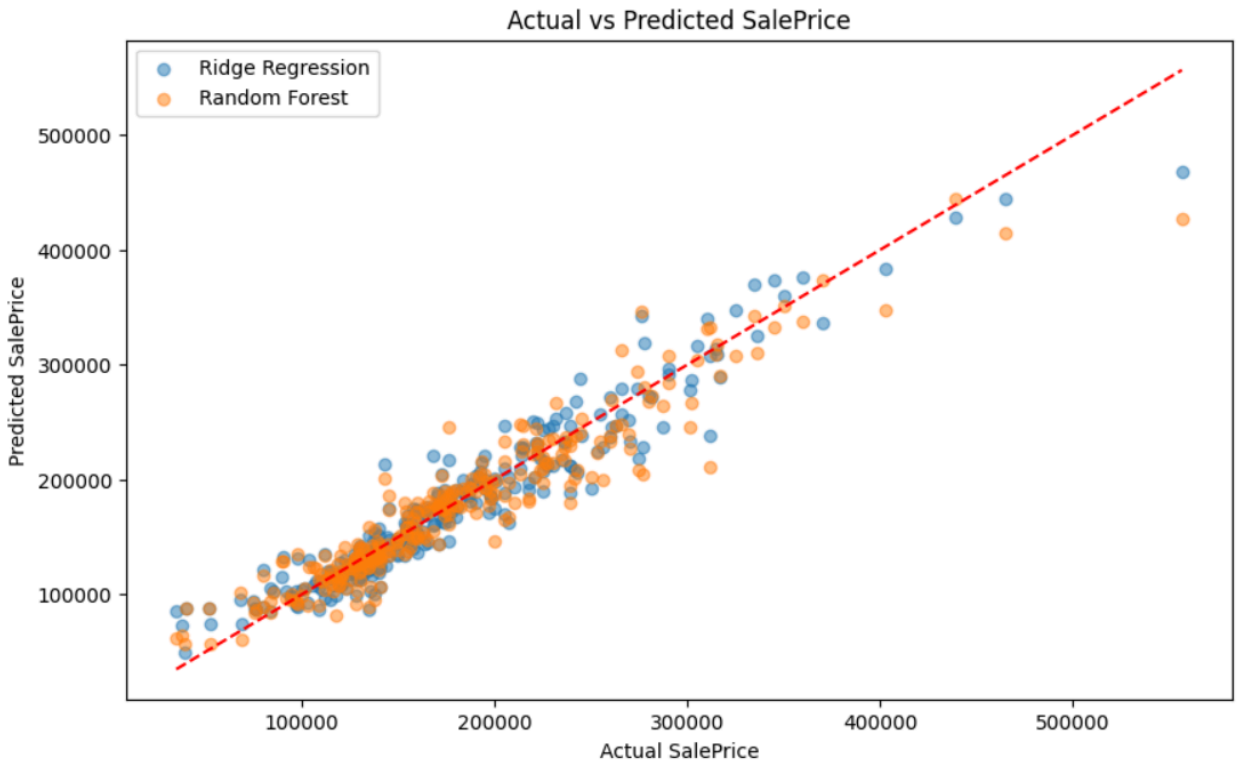
The Ridge Regression Residuals Distribution plot provides a visual representation of the difference between the actual sale prices and the predicted sale prices for the Ridge Regression model. The plot showcases a nearly normal distribution of residuals, centered around zero, indicating that the model tends to predict values close to the actual sale prices without significant skewness. However, there are still some

extreme positive and negative residuals, as indicated by the tails of the distribution. These outliers suggest that the model may struggle with certain data points, leading to prediction errors. Ideally, residuals should be randomly distributed with no clear patterns, confirming that the model fits well across various data points. The residual distribution in this case, though fairly centered, still exhibits some variance, highlighting areas where the model could be improved.



The Random Forest Residuals Distribution plot helps in visualizing how well the random forest model performed in predicting house prices. The residuals are the difference between the actual sale prices and the predicted sale prices. A well-performing model typically has residuals centered around zero with a relatively normal distribution, indicating that the errors are unbiased and that there is no systemic underprediction or overprediction.

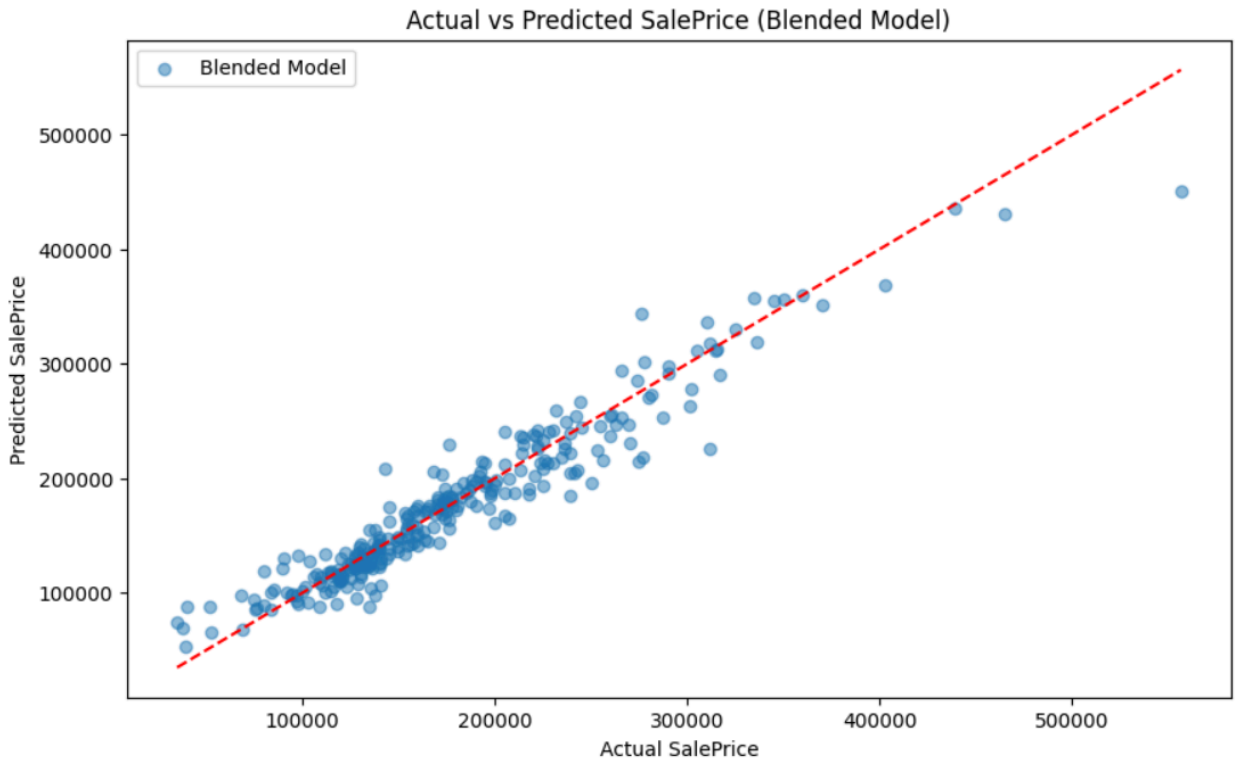
In this plot, we observe that the residuals are roughly centered around zero, which is a good sign. However, there are some noticeable tails in the distribution, suggesting that the random forest model may struggle with extreme values in the dataset, leading to larger prediction errors at both ends of the price spectrum. This suggests the model performs well for the majority of the data but may be prone to overfitting on outliers or handling extreme cases. Despite this, random forest generally provides strong predictive performance due to its ability to capture complex patterns and interactions within the data, although it may be slightly less interpretable compared to linear models like Ridge regression.



The scatter plot compares the actual versus predicted sale prices for both the Ridge and Random Forest models. Each point on the scatter plot represents a house from the test dataset, where the x-axis indicates the actual sale price and the y-axis shows the predicted price. The red dashed line represents the perfect prediction line, where the predicted sale price exactly matches the actual sale price.

Points closer to the red dashed line suggest a better model prediction, whereas points farther away from the line indicate larger prediction errors. As seen in this visualization, most points cluster along the dashed line, indicating that both the Ridge Regression and Random Forest models are producing reasonably accurate predictions. However, there are some outliers, particularly at higher price points, where predictions tend to deviate more from the actual sale prices.

This plot effectively highlights the models' strengths in predicting sale prices for a majority of houses while also revealing areas (e.g., high-value homes) where the models may underperform. It is particularly useful in visually comparing the performance of Ridge and Random Forest models in terms of accuracy across the price spectrum.



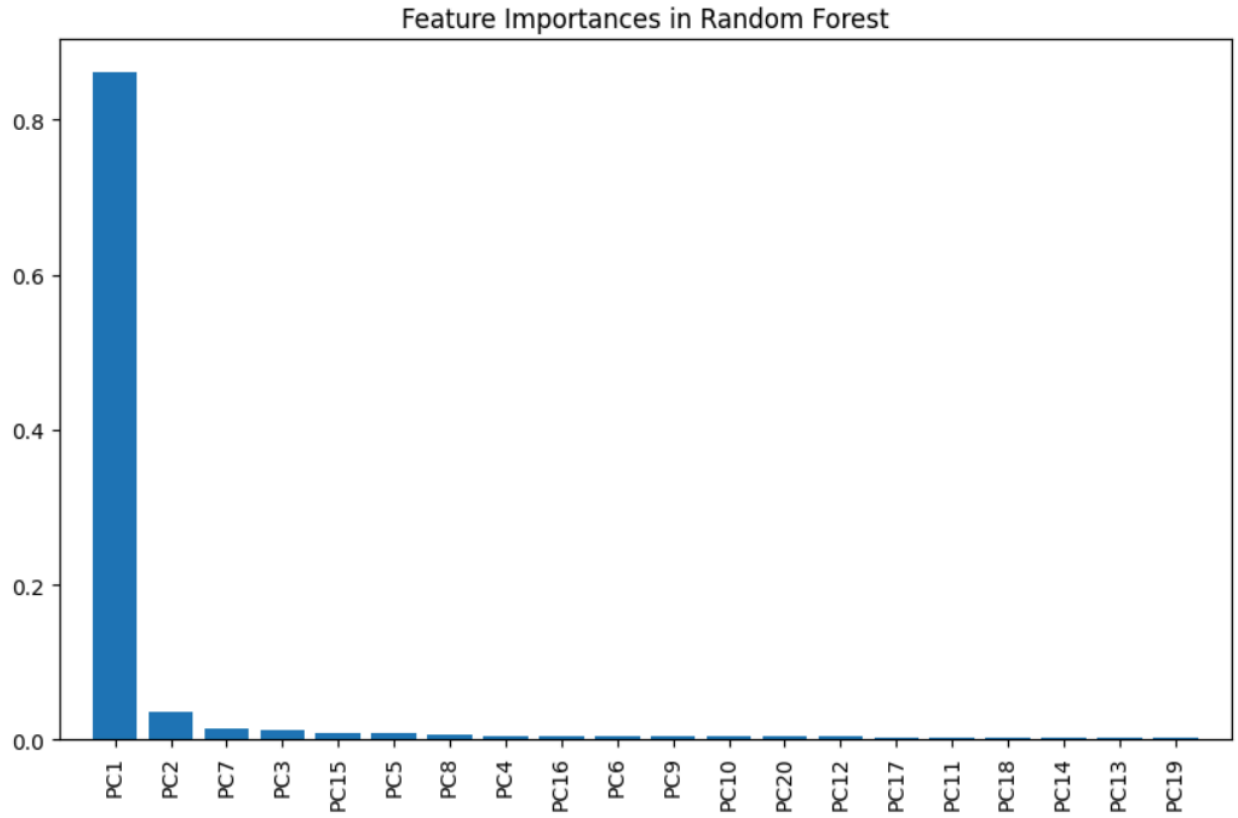
The blended model scatter plot shows the comparison between the actual sale prices and the predicted sale prices from the blended model, which combines Ridge Regression and Random Forest. Each point in the scatter plot represents a property from the test dataset, with the x-axis showing the actual sale price and the y-axis showing the predicted sale price. The red dashed line represents the perfect prediction line, where the predicted values would exactly match the actual values.

In this visualization, we can observe that most of the points cluster closely around the red line, indicating that the blended model has performed well in terms of prediction. The correlation between the actual and predicted values is strong, showing that the blended model captures the underlying patterns in the data with greater accuracy than the individual models. This plot highlights the effectiveness of blending multiple models to improve predictive performance, and the high R^2 value achieved (0.9210) supports the decision to blend the models. However, some data points show a larger deviation from the line, which may suggest areas where the model could still be improved or where outliers are present. Overall, the blended model improves the accuracy and reduces overfitting, balancing the strengths of both Ridge Regression and Random Forest.

6. Overfitting and Final Model Selection

Overfitting was a significant consideration throughout the project. While Random Forest showed strong performance, it had a tendency to overfit the training data due to the large number of trees and the model's flexibility. Ridge Regression helped mitigate overfitting by introducing regularization, but it struggled with capturing complex, non-linear relationships.

The blended model, combining Ridge Regression and Random Forest, addressed overfitting by balancing the strengths of both models. The Ridge component handled linear aspects, while Random Forest captured non-linearities without becoming overly complex. The final blended model's R-squared value of 0.9210 indicated strong predictive accuracy, with minimal overfitting risks.

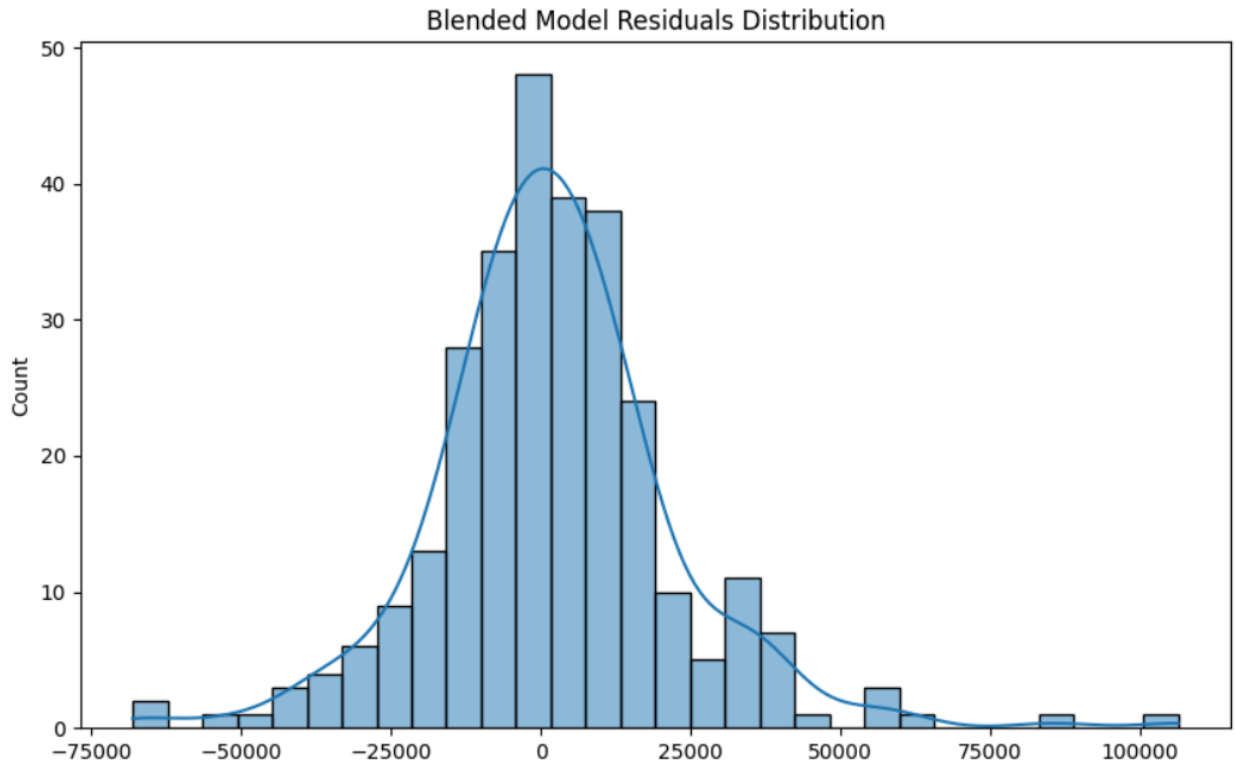


The Feature Importance plot from the Random Forest model shows how each principal component (PC) contributes to the model's prediction of house prices. The graph demonstrates that **PC1** (Principal Component 1) dominates the model's predictions with a significance value of over **0.8**, while all other components contribute far less. This indicates that PC1 contains the majority of the variance in the data, and thus has the most influence on the model's performance. PCs such as **PC2** and **PC7**, though present, have significantly smaller contributions, and the remaining components barely affect the outcome. This suggests that even though many features were used, the most important information for the model is captured primarily by PC1. This pattern is common in models that heavily rely on dimensionality reduction techniques like PCA, where the first few principal components explain most of the variance in the data.

7. Conclusion

In conclusion, this project successfully demonstrated how a combination of machine learning techniques can be used to accurately predict housing prices. By blending Ridge and Random

Forest models, a well-balanced solution that mitigates overfitting while providing reliable predictions was achieved. The importance of feature engineering, robust preprocessing, and thoughtful model selection was emphasized throughout, supported by visual insights at every stage.



The blended model residual distribution plot illustrates the distribution of residuals, which are the differences between the actual and predicted values from our final blended model. This plot is crucial in evaluating the performance of the model, as it shows how well the blended model predicts the sale prices. In an ideal scenario, residuals would be centered around zero, indicating that the model is accurate and not biased toward over- or under-predicting values.

In this case, the residuals appear to be mostly symmetric and centered around zero, which suggests that the blended model performs well in predicting sale prices. The distribution's slightly elongated tail toward positive residuals indicates some overestimation in a few instances, where the predicted values were higher than the actual values. Nonetheless, the residuals' relatively normal distribution supports the model's effectiveness.

This plot visually confirms that the blended model successfully mitigates the weaknesses of individual models, like Ridge regression and Random Forest, leading to more balanced predictions, and reducing the potential for overfitting. Overfitting occurs when a model is too closely tailored to the training data and fails to generalize well to new data, but the residual distribution suggests that the blended model strikes a balance between flexibility and accuracy.

8. Future Work

Future improvements could include:

- Experimenting with additional ensemble methods like boosting.
- Incorporating more granular location-based features, such as zip codes or proximity to amenities, which could further improve the model's accuracy.