2023

BUSINESS ANALYTICS - FINAL PROJECT

By – Gaurav Kudeshia

PROJECT GOAL

Project Goal: Predicting House Prices and Evaluating Overall Quality Using Machine Learning Models

The primary objective of this project is to develop and deploy predictive models utilizing regression and decision tree techniques. The focus is on accurately forecasting house prices based on the Zillow's Zestimate home valuation dataset (House_prices.csv). The dataset includes various variables, and the project involves selecting pertinent features as predictors for the models.

- Build a regression model capable of accurately predicting the price of a house.
- Utilize the Zillow dataset (House_prices.csv) as the primary data source.
- Identify and select relevant features from the dataset as predictors for the regression model.
- Develop a classification model to categorize houses based on the OverallQual variable.
- Define class 1 for houses with an OverallQual rating of 7 and above.
- Define class 0 for houses with an OverallQual rating below 7.
- Select relevant features as predictors for the classification model.
- Train and optimize the classification model, ensuring its ability to accurately classify houses into the defined classes.

Expected Outcome:

- Accurate regression model for predicting house prices based on selected features.
- Effective classification model for categorizing houses into OverallQual classes.
- Insights into the most influential features affecting house prices and overall quality ratings.

The project aims to provide valuable insights into the factors influencing house prices and the ability to classify houses based on their overall quality. These models can be instrumental for real estate professionals, homeowners, and potential buyers in making informed decisions.

Overview

- **Data Overview**: - The data worked on is named "House_prices" and includes variables for determining the sale price of a house. The predictor variables (independent variables) include "LotArea," "OverallQual," "YearBuilt," "YearRemodAdd," "BsmtFinSF1," "FullBath," "HalfBath," "BedroomAbvGr," "TotRmsAbvGrd," "Fireplaces," "GarageArea," and "YrSold." The sale price is the dependent variable.

  A correlation analysis was conducted to visualize relationships between variables:

  1. **Identifying Patterns:** Correlation plots helped identify patterns and relationships. The correlation matrix revealed positive or negative correlations. Strong positive correlation was observed between SalePrice (dependent variable) and OverallQual (independent variable).

  2. **Histogram Exploration:** Certain variables were explored by creating histograms.

  3. **Linear Regression Model (lm):** A linear regression model was fitted to predict SalePrice based on all variables in the "data_frame." The summary function displayed key statistics and coefficients. Significant variables were identified based on the P value, with lower P values indicating greater significance. Significant

variables included LotArea, OverallQual, YearRemodAdd, BsmtFinSF1, BedroomAbvGr, TotRmsAbvGrd, Fireplaces, and GarageArea.

4. **Comparison of Predictions:** Predicted values from the regression model were compared with actual values, showing close similarity.

5. **DecisionTree for Prediction:** DecisionTree was used to predict SalePrice based on features in the "data_frame" (Housing price) using analysis of variance ('anova'). The summary provided key statistics and information on the decision tree model and splits.

These steps provide a comprehensive overview of the analysis conducted on the house prices dataset.

After carefully observing the different parameters responsible for determining the Sales Price of a house and their descriptive statistics as shown below:

- LotArea:
  - Range: The minimum lot area is 1491 square feet, and the maximum is 215,245 square feet.
  - Distribution: The mean lot area is approximately 10,795 square feet, with a median (50th percentile) of 9,442 square feet.

- OverallQual (Overall Quality):
  - Range: The overall quality ranges from 1 (lowest) to 10 (highest).
  - Distribution: The mean overall quality is approximately 6.136, with a median of 6.000.

- YearBuilt:
  - Range: Houses in the dataset were built between 1880 and 2010.
  - Distribution: The mean year of construction is around 1971, with a median of 1973.

- YearRemodAdd (Year of Remodeling or Addition):
  - Range: Remodeling or additions occurred between 1950 and 2010.
  - Distribution: The mean year of remodeling or addition is approximately 1985, with a median of 1994.

- BsmtFinSF1 (Basement Finished Square Feet):
  - Range: Basement finished square feet range from 0 to 2260.
  - Distribution: The mean finished square feet is 446.5, and the median is 384.0.

- FullBath, HalfBath:
  - Range: Full baths range from 0 to 3, while half baths range from 0 to 2.
  - Distribution: On average, houses have around 1.564 full baths and 0.386 half baths.

- BedroomAbvGr (Bedrooms Above Ground):
  - Range: The number of bedrooms above ground ranges from 0 to 8.
  - Distribution: The mean is approximately 2.843 bedrooms, with a median of 3.000.

- TotRmsAbvGrd (Total Rooms Above Ground):
  - Range: Total rooms above ground range from 2 to 14.
  - Distribution: The mean is approximately 6.482 rooms, with a median of 6.000.

- Fireplaces:
  - Range: The number of fireplaces ranges from 0 to 3.
  - Distribution: On average, houses have around 0.628 fireplaces.

- GarageArea:
  - Range: Garage areas range from 0 to 1390 square feet.
  - Distribution: The mean garage area is 472.6 square feet, with a median of 480.0.

- YrSold:
  - Range: Houses were sold between 2006 and 2010.
  - Distribution: The mean year of sale is approximately 2008.

- SalePrice:
  - Range: Sale prices range from $34,900 to $755,000.

- Distribution: The mean sale price is $183,108, with a median of $163,000.

Comparison & Estimation of the model's performance.

Regression Model:

❖ Adjusted R-squared (R2_adj): 0.8039693

- The adjusted R-squared value of 0.8039693 indicates the proportion of the variance in the dependent variable (SalePrice) that is predictable from the independent variables in the regression model. A higher value suggests a better fit.

❖ Root Mean Squared Error (RMSE): 29381.9

- The RMSE measures the average magnitude of the errors between predicted and actual values. In this case, the RMSE of 29381.9 indicates the typical difference between the predicted and actual SalePrice values.

❖ Mean Absolute Error (MAE): 22715.48

- The MAE represents the average absolute errors between predicted and actual values. The MAE of 22715.48 indicates the average absolute difference between the predicted and actual SalePrice.

Decision Tree Model:

❖ Root Mean Squared Error (RMSE): 35295.62

- The RMSE for the decision tree model is 35295.62, suggesting the typical difference between the predicted and actual SalePrice values. It's higher than the RMSE of the regression model.

❖ Mean Absolute Error (MAE): 27271.18

- The MAE for the decision tree model is 27271.18, representing the average absolute difference between the predicted and actual SalePrice. It's also higher than the MAE of the regression model.

Comparison: The regression model generally outperforms the decision tree model in terms of both RMSE and MAE, indicating better predictive accuracy. The lower values in these metrics for the regression model suggest that it provides a more precise prediction of SalePrice compared to the decision tree model.

Insights & Conclusion

Our analysis using a regression model (lm function) reveals several key factors significantly impacting house prices. These factors, with p-values less than 0.05, indicating strong statistical significance, include:

- Lot area and overall quality: Larger lots and houses with higher overall quality ratings tend to have higher prices.
- Recent renovations: Houses with more recent renovations (higher 'YearRemodAdd' values) generally command higher prices.
- Finished basement space: Increased finished square footage in the basement ('BsmtFinSF1') positively correlates with higher prices.
- Bedrooms and rooms above ground: More bedrooms ('BedroomAbvGr') and rooms above ground ('TotRmsAbvGrd') typically lead to higher valuations.
- Fireplaces and garage size: The presence of fireplaces and larger garages also contribute to increased house value.

Correlation and price trends:

- The positive and non-zero correlation coefficients for these factors further support their influence on house prices. Notably, overall quality has the strongest influence, with a correlation coefficient of 0.7962135, indicating a nearly 80% positive association with price.

- This revised text condenses the original while maintaining key information, improving clarity, and using active voice for better readability. I hope this is helpful!

- Evaluation Metrics and Confusion Matrix:

- We utilize a confusion matrix to evaluate the performance of the classification model, incorporating key metrics like accuracy, sensitivity, specificity, and more.
- Accuracy: Achieved a rate of 83.33% in correctly classifying instances.
- Sensitivity (True Positive Rate): Demonstrates a high level of 85.45% in accurately identifying positive cases.
- Specificity (True Negative Rate): Exhibits a solid performance with a specificity of 80.00%, accurately identifying negative cases.
- Positive Predictive Value (Precision): Delivers a precision rate of 87.04% in accurately predicting positive instances.
- Negative Predictive Value: Displays a substantial predictive accuracy of 77.78% in identifying negative instances.
- Balanced Accuracy: A well-rounded metric, achieving an overall balanced accuracy rate of 82.73%.