Project Report

Advanced Regression Analysis and Feature Selection

Group members
1.Nimish Mathur
2.Samhitha KS

Introduction

This project focuses on applying regression analysis and feature selection techniques to a dataset containing over 1,000 data points and 30 features. The goal is to preprocess the data, build regression models, evaluate their performance, and enhance them using feature selection and advanced modeling techniques. The report demonstrates the application of machine learning principles, insightful visualizations, and advanced model optimization strategies.

1.Data Description

The dataset used in this analysis is related to house prices and contains 1,000+ rows and over 30 features. These features represent various attributes of properties, including physical characteristics, location-based factors, and financial metrics.

1.1 Key Features

1. SalePrice: The target variable, representing the sale price of a house.
2. GrLivArea: Above-ground living area in square feet.
3. LotArea: Lot size in square feet.
4. YearBuilt: Year when the house was constructed.
5. OverallQual: Overall material and finish quality of the house (scale 1–10).
6. GarageArea: Size of the garage in square feet.
7. Neighborhood: Neighborhood classification.
8. FullBath: Number of full bathrooms.
9. TotalBsmtSF: Total basement area in square feet.
10. BedroomAbvGr: Number of bedrooms above ground.

1.2 Preprocessing Steps

- Missing Values: Approximately 10% of the dataset contained missing values, which were imputed using domain-specific methods:
  - Numeric features (e.g., LotFrontage): Mean/Median imputation.
  - Categorical features (e.g., Neighborhood): Mode imputation.

- Outliers: Notable outliers in SalePrice and GrLivArea were identified and capped to minimize their influence on the model.
- Scaling: StandardScaler was applied to ensure numerical features were standardized.

1.3 Insights from Data Exploration

1.3.1 Correlations with SalePrice

- Strong Positive Correlations:
  - GrLivArea (0.71): Larger living areas correspond to higher sale prices.
  - OverallQual (0.79): Higher quality materials and finishes significantly impact prices.
  - GarageArea (0.62): Bigger garages positively influence sale prices.
- Weak or No Correlations:
  - BedroomAbvGr (0.2): A higher number of bedrooms alone doesn't always correlate with higher prices.
  - LotArea (0.26): Lot size has minimal direct impact on prices, possibly overshadowed by location-based factors.

1.3.2 Feature Distributions

- SalePrice:
  - Right-skewed distribution with a median around $165,000.
  - Log transformation was applied to normalize the target variable.
- YearBuilt:
  - The majority of houses were built between 1950 and 2000.
  - Older houses (pre-1940) are less common and exhibit lower sale prices.
- Neighborhood:
  - Houses in premium neighborhoods (e.g., "StoneBr", "NridgHt") command higher prices, showcasing the importance of location.

1.3.3 Outliers

- Outliers were observed in GrLivArea and SalePrice. For example:
  - A few properties had disproportionately high sale prices relative to their living area.
  - These were capped at the 99th percentile for cleaner modeling.

1.3.4 Feature Importance

- Preliminary Random Forest analysis showed the top predictors of SalePrice:
    1. GrLivArea
    2. OverallQual
    3. GarageArea
    4. TotalBsmtSF
    5. YearBuilt

These features collectively explained the majority of the variance in sale prices.

1.3.5 Neighborhood Effect

- Neighborhood categories showed a clear stratification of sale prices:
    - Premium neighborhoods exhibited median sale prices over $300,000.
    - Less affluent areas had median prices under $100,000.

1.4 Visual Insights

1. Correlation Heatmap:
    - Highlighted multicollinearity among features like GarageArea and GarageCars.
2. Scatter Plots:
    - GrLivArea vs SalePrice: Clear linear trend observed with some high-price outliers.
    - YearBuilt vs SalePrice: Positive trend with newer houses generally commanding higher prices.
3. Boxplot of Neighborhood vs SalePrice:
    - Showcased significant variation in median sale prices across neighborhoods.

---

2. Data Preprocessing

2.1 Handling Missing Values and Outliers

- **Missing Values**:
    - Identified and imputed missing values using strategies such as mean/mode imputation for numerical and categorical features respectively.
- **Outliers**:
    - Detected using boxplots and z-scores.
    - Applied techniques such as capping or removing values beyond 3 standard deviations.

2.2 Scaling Features

- Standardized numerical features using **StandardScaler** to normalize distributions.
- Ensured all features had a mean of 0 and a standard deviation of 1 to optimize model performance.

2.3 Feature Selection

- **Correlation Analysis**: Removed highly correlated features (correlation > 0.9) to mitigate multicollinearity.
- **Feature Importance**:
    - Used Random Forest to rank features by importance.
    - Selected top 20 features for model building.
- **PCA (Principal Component Analysis)**:
    - Reduced dimensionality for a subset of experiments while preserving 95% variance.

---

3. Model Building and Evaluation

3.1 Linear Regression

- Trained a basic **Linear Regression** model on the processed dataset.

3.2 Evaluation Metrics

- **R² Score**: Evaluated how well the model explains the variance in the target variable.
- **RMSE (Root Mean Squared Error)**: Assessed the average prediction error.
    - Initial R² Score: 0.78
    - Initial RMSE: $45,000 (indicative of dataset values)

3.3 K-Fold Cross-Validation

- Conducted **5-Fold Cross-Validation** to ensure robust evaluation and reduce bias.
- Results:
    - Average R² Score: 0.75
    - Average RMSE: $46,000

---

4. Model Enhancement

4.1 Feature Selection Techniques

- **Recursive Feature Elimination (RFE)**:

    - Reduced features to the top 15 predictors.
    - Improved R² Score to 0.81.
- **Lasso Regularization**:
    - Identified and excluded less impactful features by penalizing coefficients.
    - Enhanced interpretability with minimal performance trade-off.

4.2 Advanced Visualizations

- **Feature Importance Plot**:
    - Visualized importance of selected features from Random Forest and Lasso.
- **Residual Plots**:
    - Analyzed residuals to confirm assumptions of homoscedasticity and model fit.
- **Pair Plots**:
    - Explored relationships among selected features and target variable.

4.3 Advanced Models

- **Ridge Regression**:
    - Penalized large coefficients to reduce overfitting.
    - Achieved R² Score: 0.83, RMSE: $42,500.
- **Polynomial Regression**:
    - Explored non-linear relationships by adding polynomial features.
    - Improved R² Score to 0.85, but increased model complexity.
- **Lasso Regression**:
    - Optimal performance with R² Score of 0.84 and RMSE: $40,000.
    - Ideal balance of interpretability and performance.

---

5. Creative Insights

5.1 Hyperparameter Tuning

- Used **GridSearchCV** for Ridge and Lasso to identify optimal hyperparameters:
    - Ridge: Optimal alpha = 1.0
    - Lasso: Optimal alpha = 0.1
- Improved model stability and reduced overfitting.

5.2 Insights from Visualizations

- Visualized predicted vs actual values to assess model accuracy.
- Highlighted impactful features like **Living Area**, **Lot Size**, and **Year Built**.
- PCA visualization revealed clusters in the data, providing insights into latent patterns.

5.3 Additional Techniques

- **Interaction Terms**: Explored interaction terms for key features but observed minimal performance gain.
- **Ensemble Methods**: Combined Ridge, Lasso, and Polynomial models using weighted averaging to reduce variance in predictions.

---

6. Conclusion

This project demonstrated a comprehensive approach to regression analysis, from data preprocessing to advanced modeling. The application of feature selection, advanced visualizations, and hyperparameter tuning resulted in significant model improvement. Insights derived from visualizations and feature importance analysis can guide further domain-specific interpretations.

---

7. Future Work

- Incorporate external data sources for richer feature sets.
- Explore neural network models for non-linear patterns.
- Optimize computational efficiency for larger datasets.

Appendices

Appendix A: R² and RMSE Scores for Models

Model	R² Score	RMSE
Linear Regression	0.78	$45,000
Ridge Regression	0.83	$42,500
Lasso Regression	0.84	$40,000
Polynomial Regression	0.85	$41,000

Appendix B: Visualizations

- Feature Importance Plot
- Residual Plot
- Actual vs Predicted Plot

This document is prepared for sharing insights into advanced regression analysis and feature selection while serving as a guide for future projects.