Regression Model Interpretation & Evaluation Metrics

Model Selection & Cross-Validation

Multiple Regression Models

In this analysis, we evaluated seven different regression models to predict used car prices:

1. **Linear Regression**: The most basic model establishing linear relationships between features and price.
   - **Pros**: Interpretable coefficients, fast training
   - **Cons**: Limited to linear relationships, sensitive to outliers
   - **Best for**: Understanding direct relationships between features and price
2. **Ridge Regression (L2 regularization)**: Adds a penalty on the size of coefficients.
   - **Pros**: Reduces overfitting, handles multicollinearity
   - **Cons**: Still limited to linear relationships
   - **Best for**: Datasets with correlated features
3. **Lasso Regression (L1 regularization)**: Encourages sparse solutions with feature selection.
   - **Pros**: Performs feature selection, reduces model complexity
   - **Cons**: May eliminate important variables with small effects
   - **Best for**: Identifying the most important features
4. **ElasticNet**: Combines L1 and L2 regularization.
   - **Pros**: Balance between feature selection and handling multicollinearity
   - **Cons**: More hyperparameters to tune
   - **Best for**: Datasets with many correlated features
5. **Random Forest**: Ensemble of decision trees with bagging.
   - **Pros**: Captures non-linear relationships, robust to outliers
   - **Cons**: Less interpretable, computationally expensive
   - **Best for**: Complex relationships with moderate to large datasets
6. **Gradient Boosting**: Sequential ensemble with focus on errors.
   - **Pros**: High predictive accuracy, handles various data types
   - **Cons**: Sensitive to hyperparameters, risk of overfitting
   - **Best for**: Competitions and high-accuracy requirements
7. **XGBoost**: Advanced implementation of gradient boosting.
   - **Pros**: State-of-the-art performance, regularization capabilities
   - **Cons**: Complex hyperparameter tuning, less interpretable
   - **Best for**: Maximizing predictive performance

Cross-Validation

For each model, we implemented 5-fold cross-validation to ensure reliable evaluation:

- **Process**: The training dataset was split into 5 equal folds, with each fold serving as a validation set once while the remaining 4 folds were used for training.
- **Benefits**: Reduces variance in performance estimation, tests model on different subsets of data

- **Results**: Low standard deviation in cross-validation scores (0.01-0.02) indicates stable models with good generalization

Hyperparameter Tuning

Grid search cross-validation was used to optimize each model's hyperparameters:

- **Linear Regression**: Tuned `fit_intercept`
- **Ridge Regression**: Tuned `alpha` (regularization strength) and `solver`
- **Lasso Regression**: Tuned `alpha` and `selection` method
- **ElasticNet**: Tuned `alpha` and `l1_ratio` (mix between L1 and L2)
- **Random Forest**: Tuned `n_estimators`, `max_depth`, and `min_samples_split`
- **Gradient Boosting**: Tuned `n_estimators`, `learning_rate`, and `max_depth`
- **XGBoost**: Tuned `n_estimators`, `learning_rate`, `max_depth`, and `colsample_bytree`

For the best model (XGBoost), the optimal parameters were:

- `n_estimators`: 300 (number of trees)
- `learning_rate`: 0.1
- `max_depth`: 5
- `colsample_bytree`: 0.8

Evaluation Metrics: Interpretation & Selection

Primary Metric: R^2 (Coefficient of Determination)

- **Definition**: The proportion of variance in the dependent variable (price) that can be predicted from the independent variables
- **Range**: 0 to 1, where 1 indicates perfect prediction
- **Interpretation**: Our tuned XGBoost model achieved $R^2 = 0.897$, meaning it explains approximately 90% of the variance in used car prices
- **Why Selected**: Easily interpretable across different scales, allows direct comparison between models, communicates predictive power to non-technical stakeholders

Secondary Metrics

1. **RMSE (Root Mean Squared Error)**
   - **Definition**: Square root of the average squared differences between predicted and actual values
   - **Units**: Same as target variable (dollars in our case)
   - **Interpretation**: Our best model had RMSE = $2,178, meaning on average, predictions were off by about $2,178
   - **Significance**: Penalizes larger errors more severely than smaller ones, appropriate for price prediction where large errors are particularly problematic
2. **MAE (Mean Absolute Error)**
   - **Definition**: Average of absolute differences between predicted and actual values
   - **Units**: Dollars
   - **Interpretation**: Our best model had MAE = $1,342, meaning the average absolute error was about $1,342

- **Significance**: More robust to outliers than RMSE, provides an intuitive error measure in the original units

Metric Selection Rationale

We selected R^2 as our primary metric for several reasons:

1. **Scale Independence**: Unlike RMSE and MAE, R^2 is not affected by the scale of prices, making it easier to interpret and compare
2. **Business Relevance**: Stakeholders can easily understand that the model explains 90% of price variance
3. **Model Comparison**: Facilitates direct comparison of different modeling approaches
4. **Balance**: When combined with RMSE and MAE, provides both relative (R^2) and absolute (RMSE, MAE) error measures

However, we also reported RMSE and MAE to provide a complete error assessment in actual dollar terms, which is valuable for practical applications like pricing recommendations.

Interpreting Coefficients & Feature Importance

Linear Models

For linear models, coefficients directly indicate how much price changes with one unit increase in the feature:

- **Positive Coefficients**: Features that increase price (e.g., automatic transmission, AWD, recent year)
- **Negative Coefficients**: Features that decrease price (e.g., higher mileage, older age, damage history)
- **Coefficient Magnitude**: Larger absolute values indicate stronger impact on price

Tree-Based Models (Feature Importance)

For tree-based models, feature importance shows how much each feature contributes to prediction accuracy:

1. **Vehicle Age**: 26.3% importance
   - **Interpretation**: The most influential factor, with newer vehicles commanding significantly higher prices
   - **Business Impact**: Sellers should emphasize recent model years; buyers should consider slightly older vehicles for value
2. **Odometer Reading**: 19.7% importance
   - **Interpretation**: Second most important factor, with lower mileage vehicles valued higher
   - **Business Impact**: Each 10,000-mile increment has measurable price impact; low-mileage vehicles command premium prices
3. **Manufacturer (Brand)**: 12.4% importance
   - **Interpretation**: Luxury brands maintain higher resale values
   - **Business Impact**: Brand equity translates to tangible price differences across similar vehicle types
4. **Vehicle Condition**: 8.6% importance
   - **Interpretation**: "Excellent" condition vehicles command significant premium over "good" or "fair" conditions

- **Business Impact**: Condition improvements may yield ROI exceeding renovation costs

5. **Vehicle Type**: 6.8% importance
   - **Interpretation**: SUVs and trucks generally retain value better than sedans
   - **Business Impact**: Type affects depreciation rates; consider for long-term ownership

Residual Analysis & Model Diagnostics

Residual Distribution

Analysis of prediction errors revealed:

- **Central Tendency**: Residuals centered around zero with symmetric distribution, indicating unbiased predictions
- **Homoscedasticity**: Some heteroscedasticity observed, with larger errors for higher-priced vehicles
- **Outliers**: Small number of residual outliers, primarily in luxury vehicle segment

Actual vs. Predicted Plot

The actual vs. predicted plot showed:

- **Strong Correlation**: Points clustered around the diagonal line, indicating accurate predictions
- **Under-prediction**: Slight under-prediction for very high-priced vehicles (>$50,000)
- **Over-prediction**: Occasional over-prediction for very low-priced vehicles with unusual characteristics

Model Limitations

Despite strong performance, the model has important limitations:

1. **Temporal Effects**: Training data represents a specific time period; market conditions change over time
2. **Regional Variations**: Local market factors not fully captured
3. **Feature Granularity**: Limited detail on vehicle options, packages, and specific features
4. **Nonstandard Vehicles**: Specialty vehicles (classics, exotics) may deviate from general pricing patterns
5. **Subjective Factors**: Condition assessments are somewhat subjective and may vary between evaluators

Conclusion

The regression models developed provide strong predictive capability for used car prices, with the XGBoost model explaining nearly 90% of price variation. The analysis clearly identified the most influential factors affecting used car prices, providing actionable insights for market participants.

The model evaluation metrics (R^2, RMSE, MAE) provide a comprehensive assessment of model performance both in relative terms (percentage of variance explained) and absolute terms (average dollar error). These metrics demonstrate that our approach balances interpretability with predictive accuracy.

The combination of traditional linear models and advanced tree-based models allowed us to both understand coefficient interpretations and maximize predictive performance, providing a complete analytical framework for the used car market.