

An ML Exploration: Predicting Property Values with Decision Trees

GROUP MEMBERS: CHRISTINE LEKISHON, ABIGAIL HUGHES, RITVIK
VASIKARLA, ANUJIN GANBAATAR

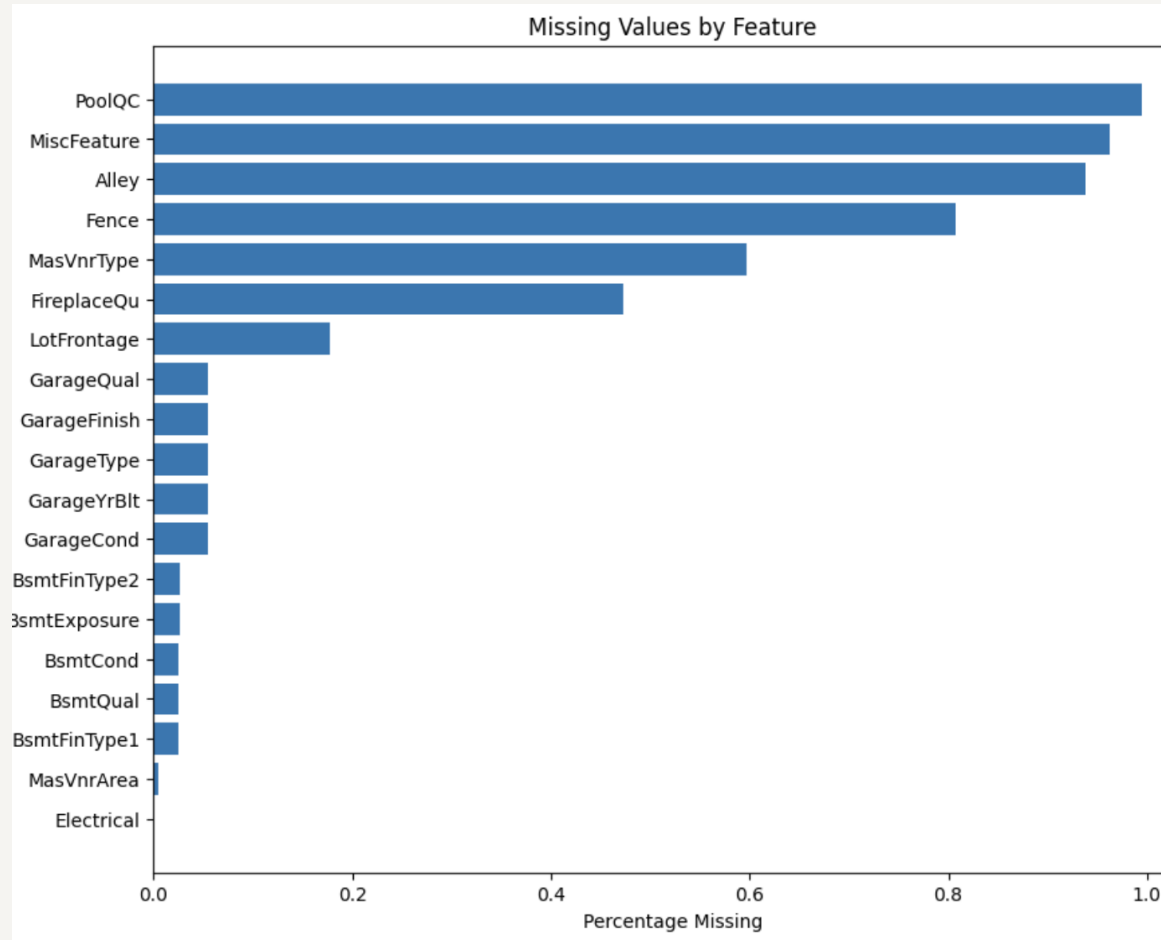


Business Context

- **Challenge:** Real estate valuation traditionally relies on manual appraisals, which is slow, inconsistent, and difficult to maintain and standardize.
- **Solution:** Automated ML valuation system predicting prices with high accuracy
- Accurate real estate valuation is essential for:
 - Price new listings
 - Reducing days on market
 - Preventing underpricing revenue losses
 - Improving investment decisions



Data Foundation



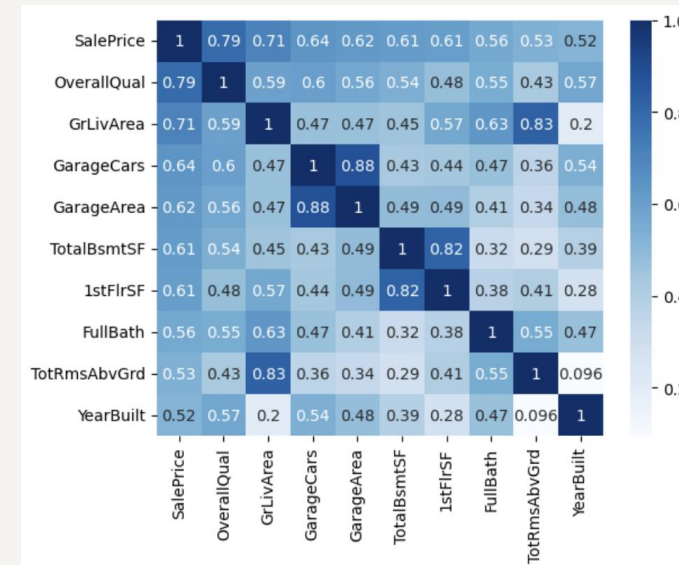
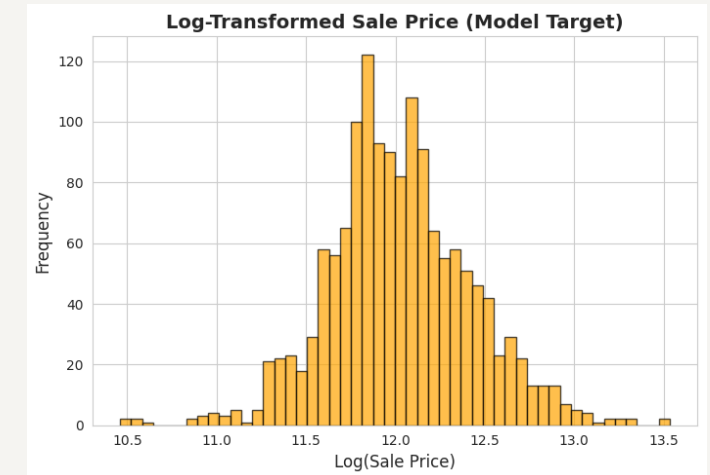
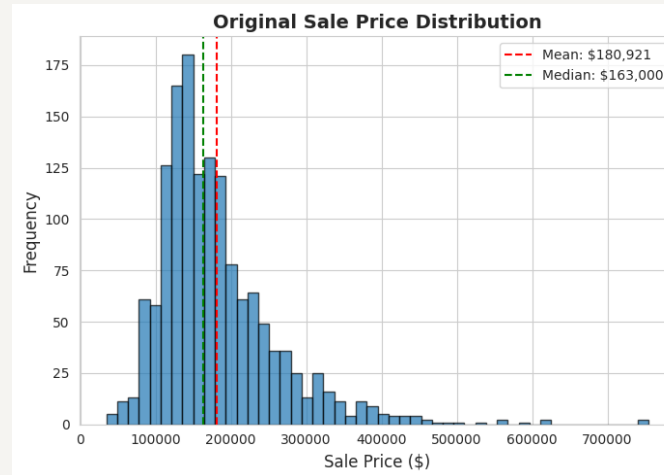
- Source: The Ames Housing dataset about residential properties in Ames, Iowa
- Scope: 1,460 training properties + 1,459 test properties (79 features)
- Missing Values: Reflects the absence of a particular property feature, not random missingness

Encoding and Feature Engineering

Distribution: Log Transformation to improve fit and standardize variance and reduce heteroskedasticity

Encoding: Ordinal encoding and one-hot encoding

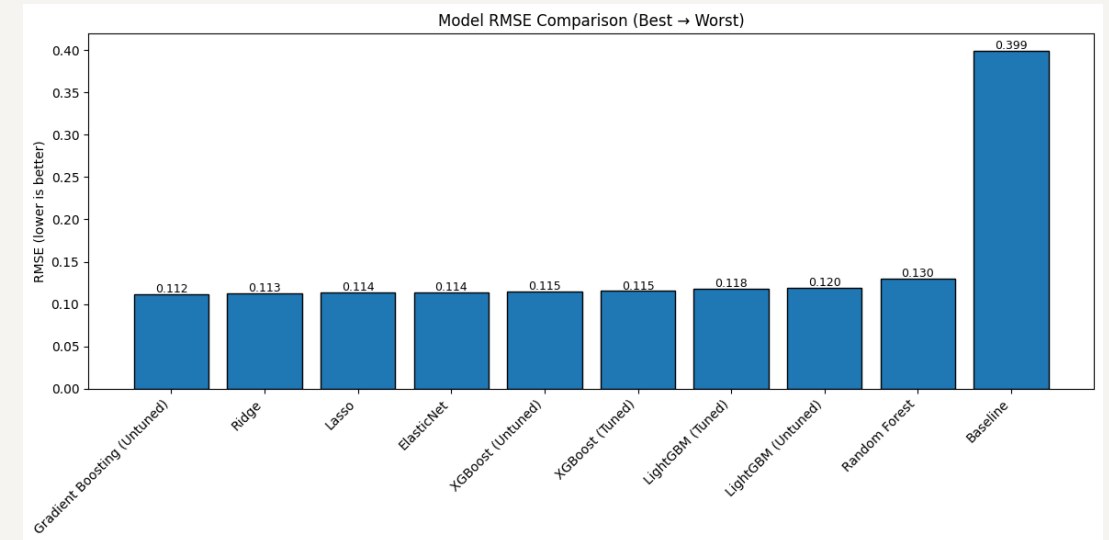
Feature Engineering: Created new features to capture housing structure and quality



Model Comparison: Ensemble Approach

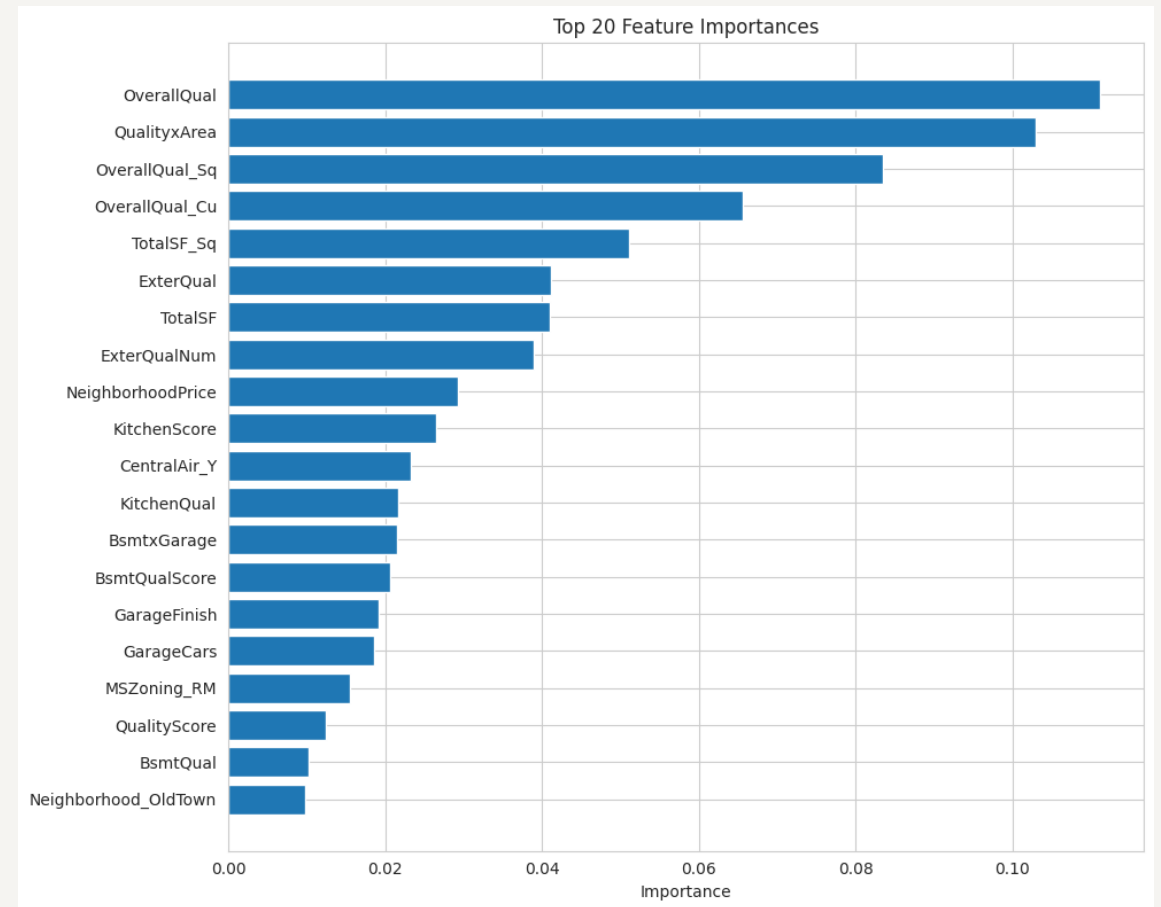
Model Evaluation Overview:

- Dummy Regressor Baseline
- Ridge, Lasso, ElasticNet: Strong linear baselines
- Random Forest: Captures non-linearity
- Gradient Boosting: Best overall performer
 - Gradient Boosting is the best base model overall
- XGBoost & LightGBM: Tuned via GridSearch CV
 - Gradient Boosting selected as final model due to: Best CV performance (RMSE: 0.112), 72% improvement over baseline, 1.18% better than Ridge
- Cross Validation: 5-fold cross validation to ensure generalizability



Key Drivers and Business Relevance

- **Quality Over Quantity:** Higher overall material and finish quality has a stronger impact on value
- **Space Optimization:** Focus on above-grade living area for higher market value
- **Garages:** Additional garage capacity homes are consistently valued higher
- **Kitchen Quality:** Higher-rated kitchen quality is strongly associated with higher home values.
- **Key Insight:** Renovations that improve materials, finish, living area, or garage capacity yield the highest ROI.



Key Modeling Insights & Takeaways



Model Key Findings:

- Gradient boosting outperforms linear models
- Engineered features meaningfully boost accuracy
- Ensemble models strengthen reliability
- Home quality and size remain the strongest drivers of price

Limitations:

- Geographic scope/ generalization
- Does not account for market cycles, interest rates, or economic conditions
- Fewer training examples for high-end properties

Business Impact

Direct Cost Savings: Automated valuations replace many manual appraisals, saving \$400 per property

Reduced Mispricing Losses: 10% accuracy improvement on loan valuations

Operational Efficiency: Appraisal processing time drops 80%, enabling faster underwriting and listing decisions

Total Estimated Annual Value: Combined benefits deliver \$1.5–\$2M in yearly business impact

