House Price Prediction Project Report

Overview

This project aimed to predict house prices (log(SalePrice)) for the Ames Housing dataset, initially targeting an R^2 of 95%, later adjusting to match a top Kaggle competitor's RMSE of 0.1071 (log scale). The best model was a stacked ensemble of XGBoost and LightGBM with Recursive Feature Elimination (RFE) to select the top 50 features.

Methodology

- **Data:** 1,168 training samples, 292 test samples, 238 features (after preprocessing).
- **Feature Engineering:**
  - Created `TotalSF` (sum of `TotalBsmtSF`, `1stFlrSF`, `2ndFlrSF`), `HouseAge` (`YrSold - YearBuilt`), `OverallQual_SF` (`OverallQual * TotalSF`), and `OverallQual_GrLivArea` (`OverallQual * GrLivArea`).
  - Added polynomial features: `GrLivArea_Squared`, `TotalSF_Squared`.
  - Added interaction terms: `Neighborhood_*_OverallQual` (e.g., `Neighborhood_Crawfor_OverallQual`), and later `OverallQual_YearBuilt`, `OverallQual_GarageArea`.
  - Added `TotalBathrooms` (sum of full and half bathrooms, weighted) in later iterations.
  - Added spatial clustering: `Neighborhood_Cluster` using K-means in later iterations.
  - Applied feature capping: `GrLivArea` at 5,500 (later 6,000), `TotalSF` at 11,000 (later 12,000), `OverallQual_SF` at 70,000 (later 80,000), `OverallQual_GrLivArea` at 70,000 (later 80,000).
- **Preprocessing:**
  - Handled missing values: numerical with median, categorical with mode.
  - Encoded categorical variables using one-hot encoding.
  - Scaled features using `StandardScaler`.
- **Models Tried:**
  - Linear Regression with RFE (50 features): Initial model, validation RMSE 0.1226.
  - Elastic Net, XGBoost, Gradient Boosting, weighted ensembles, and stacking.
  - Best model: Stacked ensemble of XGBoost and LightGBM with RFE-selected features, using a LinearRegression meta-model.
  - Later iterations: Added CatBoost and used a GradientBoosting meta-model, but these regressed performance.
- **Evaluation:** RMSE and R^2 on log(SalePrice).

Best Result

- **Initial Model (Linear Regression with RFE):**
  - Validation RMSE (on test set): 0.1226 (log scale), R^2: 0.9185.
  - Cross-Validation RMSE: 0.1449 (log scale), R^2: 0.8645.
  - Kaggle Score: 0.18241 (log scale), top 40-50%.
- **Intermediate Model (XGBoost with RFE):**
  - Cross-Validation RMSE: 0.1325 (log scale), R^2: 0.8894.
  - Kaggle Score: 0.13195 (log scale), top 15-20%.
- **Intermediate Model (XGBoost + LightGBM Ensemble):**
  - Cross-Validation RMSE (XGBoost): 0.1259 (log scale), R^2: 0.9002.
  - Cross-Validation RMSE (LightGBM): 0.1345 (log scale), R^2: 0.8858.
  - Kaggle Score: 0.124 (log scale, estimated), top 10-15%.
- **Best Model (XGBoost + LightGBM Stacking):**
  - Cross-Validation RMSE (XGBoost): 0.1216 (log scale), R^2: 0.9069.
  - Cross-Validation RMSE (LightGBM): 0.1253 (log scale), R^2: 0.9009.
  - Predictions: $125,750 to $189,393 (first few rows), full range $37,600 to $613,000 (log(SalePrice) from 10.5296 to 13.3260).
  - Kaggle Score: 0.12565 (log scale), rank 659, top 10-15%.
- **Intermediate Model (XGBoost + LightGBM + CatBoost Stacking):**
  - Features: `TotalSF`, `HouseAge`, `OverallQual_SF`, `OverallQual_GrLivArea`, `GrLivArea_Squared`, `TotalSF_Squared`, `OverallQual_YearBuilt`, `Neighborhood_Cluster`, and others selected by RFE (e.g., `Neighborhood_Crawfor_OverallQual`).
  - Cross-Validation RMSE (XGBoost): 0.1222 (log scale), R^2: 0.9061.
  - Cross-Validation RMSE (LightGBM): 0.1257 (log scale), R^2: 0.9003.
  - Cross-Validation RMSE (CatBoost): 0.1200 (log scale), R^2: 0.9093.
  - Predictions: $125,172 to $190,184 (first few rows), full range $35,500 to $590,000 (log(SalePrice) from 10.475994035951858 to 13.286360957271917).
  - Kaggle Score: 0.12728 (log scale), top 10-15%.
- **Final Model (XGBoost + LightGBM + CatBoost Stacking with GradientBoosting Meta-Model):**
  - Added Features: `TotalBathrooms`, `OverallQual_GarageArea`.
  - Cross-Validation RMSE (XGBoost): 0.1198 (log scale), R^2: 0.9082.
  - Cross-Validation RMSE (LightGBM): 0.1237 (log scale), R^2: 0.9023.
  - Cross-Validation RMSE (CatBoost): 0.1176 (log scale), R^2: 0.9121.
  - Predictions: $129,953 to $184,170 (first few rows), full range $36,000 to $550,000 (log(SalePrice) from 10.487165058268593 to 13.225459713470128).
  - Kaggle Score: 0.13257 (log scale), top 15-20%.

Kaggle Submission

- **Initial Submission (Linear Regression):**

  - Prediction Range: $30,233 to $50,769 (first few rows, log(SalePrice) from 9.339 to 12.813).
  - Kaggle Public Score: 0.18241 (log scale), top 40-50%.
- **Intermediate Submission (XGBoost):**
  - Prediction Range: $137,534 to $193,474 (first few rows), full range $45,300 to $514,000.
  - Kaggle Public Score: 0.13195 (log scale), top 15-20%.
- **Intermediate Submission (XGBoost + LightGBM Ensemble):**
  - Prediction Range: $122,458 to $191,414 (first few rows), full range $48,600 to $494,000.
  - Kaggle Public Score: 0.124 (log scale, estimated), top 10-15%.
- **Best Submission (XGBoost + LightGBM Stacking):**
  - Prediction Range: $125,750 to $189,393 (first few rows), full range $37,600 to $613,000.
  - Kaggle Public Score: 0.12565 (log scale), rank 659, top 10-15%.
- **Intermediate Submission (XGBoost + LightGBM + CatBoost Stacking):**
  - Prediction Range: $125,172 to $190,184 (first few rows), full range $35,500 to $590,000.
  - Kaggle Public Score: 0.12728 (log scale), top 10-15%.
- **Final Submission (XGBoost + LightGBM + CatBoost Stacking with GradientBoosting Meta-Model):**
  - Prediction Range: $129,953 to $184,170 (first few rows), full range $36,000 to $550,000.
  - Kaggle Public Score: 0.13257 (log scale), top 15-20%.

Comparison with Target

- **Target (Mubashir Qasim's Team):** RMSE 0.1071 (log scale), estimated R^2 ~0.9282.
- **Gap (Initial Submission):** ~70% improvement needed (0.18241 vs. 0.1071).
- **Gap (Intermediate Submission, XGBoost):** ~23% improvement needed (0.13195 vs. 0.1071).
- **Gap (Intermediate Submission, Ensemble):** ~16% improvement needed (0.124 vs. 0.1071).
- **Gap (Best Submission, XGBoost + LightGBM Stacking):** ~17.4% improvement needed (0.12565 vs. 0.1071); ~8.4% to top 10% (0.12565 vs. 0.115).
- **Gap (Intermediate Submission, XGBoost + LightGBM + CatBoost Stacking):** ~18.8% improvement needed (0.12728 vs. 0.1071); ~9.7% to top 10% (0.12728 vs. 0.115).
- **Gap (Final Submission):** ~23.8% improvement needed (0.13257 vs. 0.1071); ~13.3% to top 10% (0.13257 vs. 0.115).
- **Competitiveness:** The best score of 0.12565 (rank 659, top 10-15%) is competitive but did not reach the top 10% (0.115). Later iterations with

CatBoost and a GradientBoosting meta-model regressed performance due to overfitting and underprediction of high-value houses.

Visualizations

- **Actual vs Predicted:** Shows tight alignment with actual values (R^2 0.9185).
- **Residual Plot:** Residuals mostly random, with some spread.
- **Feature Importance:** Top features include `OverallQual`, `GrLivArea`, `TotalSF`.

Challenges

- The initial linear regression model underpredicted house prices on the Kaggle test set (predictions $30,233 to $50,769), leading to a high RMSE (0.18241).
- Feature capping limited the model's ability to predict high-value houses in earlier iterations.
- The Kaggle test set likely has a different distribution of house prices, causing underpredictions in the initial model.
- Overfitting to the validation set resulted in an overly optimistic validation RMSE (0.1226).
- The addition of CatBoost and new features led to a regression in Kaggle score (0.12565 to 0.12728), likely due to overfitting and underprediction of high-value houses.
- The final model with a GradientBoosting meta-model further regressed (0.13257), likely due to overfitting, increased variability in cross-validation, and suboptimal meta-model performance.

Conclusion

The initial linear regression model achieved a validation RMSE of 0.1226 but did not generalize well to the Kaggle test set (RMSE 0.18241). Switching to XGBoost improved the score to 0.13195, and further enhancements (hyperparameter tuning, advanced feature engineering, and ensembling with LightGBM) achieved a score of 0.124 (estimated). The best model, a stacked ensemble of XGBoost and LightGBM with a LinearRegression meta-model, achieved a Kaggle score of 0.12565 (rank 659, top 10-15%), missing the top 10% target (~0.115) by ~8.4%. Later iterations with CatBoost, new features (`TotalBathrooms`, `OverallQual_GarageArea`), and a GradientBoosting meta-model regressed performance to 0.12728 and 0.13257, respectively, due to overfitting and underprediction of high-value houses. This project demonstrates the importance of iterative model improvement, feature engineering, and advanced ensembling techniques, as well as the challenges of generalizing to a test set with a different distribution. The final score of 0.12565 is a strong result for a portfolio project, showcasing skills in machine learning, feature engineering, and model ensembling.

Files

- `final_model_xgboost_rfe.pkl`: Saved XGBoost model.
- `final_model_lightgbm_rfe.pkl`: Saved LightGBM model.
- `final_model_catboost_rfe.pkl`: Saved CatBoost model (from final iteration).
- `submission.csv`: Best Kaggle submission file (score 0.12565).
- `submission-4.csv`: Intermediate Kaggle submission file (score 0.12728).
- `submission-5.csv`: Final Kaggle submission file (score 0.13257).
- `selected_features.pkl`: RFE-selected features.
- `scaler.pkl`: Fitted scaler.
- `plots/`: Contains visualizations (`actual_vs_predicted.png`, `residual_plot.png`, `feature_importance.png`).
- `code/final_script.py`: Standalone script to reproduce the results.