**Model Card: House Price Prediction**

Basic Information

* **Author**: Chipo Jaya, chipo.jaya@gwu.edu
* **Date**: 12/09/2024
* **Model Version**: 1.0
* **License**: MIT License
* **Description:** This project aims to analyze the Ames Housing dataset and build a model to predict house prices based on various features. The goal is to understand the factors influencing house prices and develop an effective predictive model.

Intended Use

* **Primary Use Case**: Predicting sale prices of residential properties based on features such as location, quality, and size.
* **Target Audience:** Real estate professionals, home buyers, and investors seeking to understand property values.

Training and Test Data

* **Source:** Kaggle’s *House Prices - Advanced Regression Techniques* competition dataset.
* **Rows in Training Data:** 1,460
* **Rows in Test Data:** 1,459 (without SalePrice).

Key Features:

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Modeling Role | Measurement Level | Description |
| *SalePrice* | Target | Continuous | The property's sale price in dollars |
| *LotArea* | Feature | Continuous | Lot size in square feet |
| *OverallQual* | Feature | Ordinal | Overall material and finish quality |
| *YearBuilt* | Feature | Discrete | Original construction date |
| *Neighborhood* | Feature | Discrete | Location-based categorization |
| *GrLivArea* | Feature | Continous | Above ground living area in square feet |
| *YearBuilt* | Feature | Discrete | Original construction date |

Model Details

* **Target Column**: SalePrice
* **Type of Model**: XGBoost Regressor
* **Software**: Google Collab, scikit-learn 0.24.2, XGBoost 1.4.2
* **Hyperparameters:**
  1. *n\_estimators*: 200
  2. *max\_depth*: 3
  3. *learning\_rate*: 0.1

Evaluation Metrics

* **Metrics Used**:
  + Root Mean Squared Error (RMSE)
  + R-squared (R²)

Performance Summary

* **Best Model:** XGBoost achieved the best performance with:
  + RMSE: **31178.57** (Cross-validation average)
  + R²: **0.9998** (Training set)

Visualizations

**Housing Types:**

A pie chart with numbers and a pie chart

Description automatically generated

**Distribution of Sale Prices**:

A graph of a distribution of sales

Description automatically generated

**A graph of a bar graph

Description automatically generated with medium confidenceOver Quality:**

**A graph of a graph with numbers and a chart

Description automatically generated with medium confidenceYear Built**

**Heatmap of Key Variables**

A screenshot of a graph

Description automatically generated

**Key Insights**

* High correlation between OverallQual and SalePrice underscores the importance of material quality in property valuation.
* LotArea and Neighborhood contribute to SalePrice but with significant variability.
* XGBoost excels at capturing complex, non-linear relationships compared to traditional linear models.

Ethical Considerations

**Potential Negative Impacts**

* **Math or Software Problems**:
  + Potential overfitting to training data
  + Bias towards features more common in the training dataset
* **Real-world Risks**:
  + May perpetuate existing biases in housing markets
  + Could be misused to justify discriminatory pricing practices

Uncertainties

* **Math or Software Problems**:
  + Model performance on data significantly different from the training set
  + Handling extreme outliers or unusual property characteristics
* **Real-world Risks**:
  + Changes in market conditions not reflected in the training data
  + Regional differences in housing markets not captured by the model

[Link to my notebook] <https://github.com/chipojaya1/housing-prices-prediction/blob/main/house_prices_model.ipynb>