House Price Prediction using Regression

Course Project Report

Landon Magana

Abstract

This project applies machine learning technique to the problem of predicting residential house sale prices from property features. I decided to got with an application based machine learning study using the Ames housing dataset, which contains property sales from Ames, Iowa and includes numerical and categorical variables such as living area, year built, garage capacity, and other housing characteristics. This project shows how regression is useful for real estate price estimation and provides a practical exampled of applying machine learning to real world problems.

Introduction

The selected topic for this course project is regression. Regression is a supervised machine learning approach used when the goal is to predict a continuous numeric value instead of a category (Scikit-learn Developers, n.d.). In this project, the continuous value is the selling price of a house. House price prediction is useful because buyers, sellers, banks, real estate agents, and city planners often need an estimate of property value. Manual price estimates can be difficult because a home price depends on many factors at the same time, including size, location, age, and many other factors.

The main problem addressed in this paper is how regression can be used to estimate residential house prices from structured housing data. The machine learning model learns patterns from past home sales and then predicts the sale price for homes it has not seen before. This makes the problem a standard supervised learning task.

Related work

Zhong explains that regression is used when the goal is to predict a continuous output variable, usually written as y, based on one or more predictor variables, usually written as x (Zhong, n.d.). In the context of this project, the predictor variables are housing features such as square footage, number of bedrooms, and location, while the output variable is the predicted house sale price.

Another issue in regression is model fitting. A model that is too simple may underfit the data, while a model that is too complex may overfit the training data and perform poorly on unseen testing data. Zhong explains that model selection should focus on minimizing error on testing data rather than only improving training performance (Zhong, n.d.). Regularization is one method used to reduce overfitting in regression models. Zhong explains that regularization adds a penalty term to the loss function so that the model's weight stays smaller.

Methodology

The methodology follows a standard supervised regression workflow. First, the dataset is loaded and the target variable, salePrice, is separated from the input features. Next, unnecessary identifier columns are removed because they do not describe the actual property. The remaining features are divided into numerical and categorical variables.

Missing values are handled differently depending on feature type. For numerical features, missing values are replaced using the median value from the training data. For categorical features, missing values are replaced using the most frequent category.

After preprocessing, the data is split into a training set and a testing set. The training set is used to learn the model parameters, and the testing set is held out until the final evaluation.

The Keras model is a feedforward neural network. The input layer receives the processed housing features. The hidden layers use rectified linear unit activation functions, which allow the model to learn nonlinear relationships.

Experiment

The experiment uses the Ames Housing dataset. The output files include a training history plot, an actual-versus-predicted scatter plot, a residual plot, a metrics CSV file, and a sample predictions CSV file. The actual-versus-predicted plot is used to visually evaluate the model. If the model performs well, most points should appear close to the diagonal line. The residual plot is also useful. Residuals are calculated as actual price minus predicted price. A good model should have residuals scattered around zero.



Figure 1: Acual vs Predicted House Prices

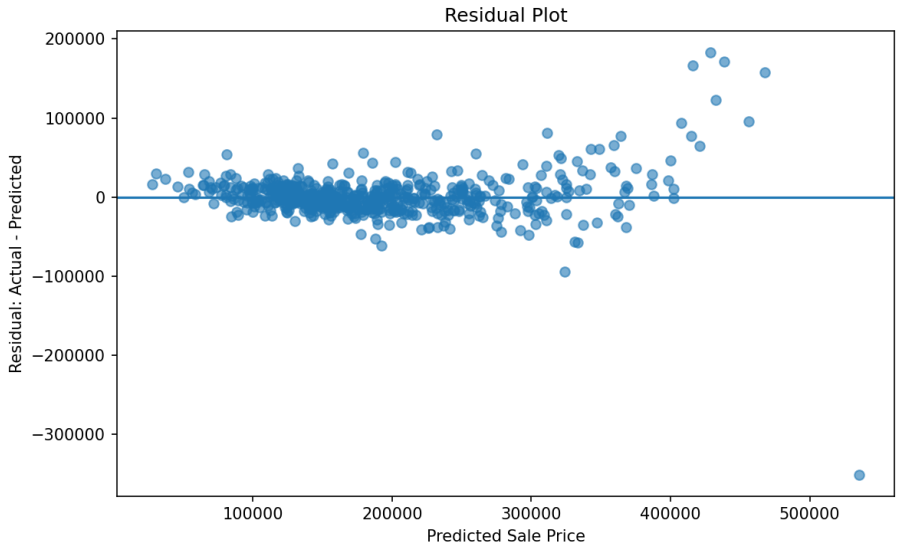


Figure 2: Residual Plot for House price Predictions

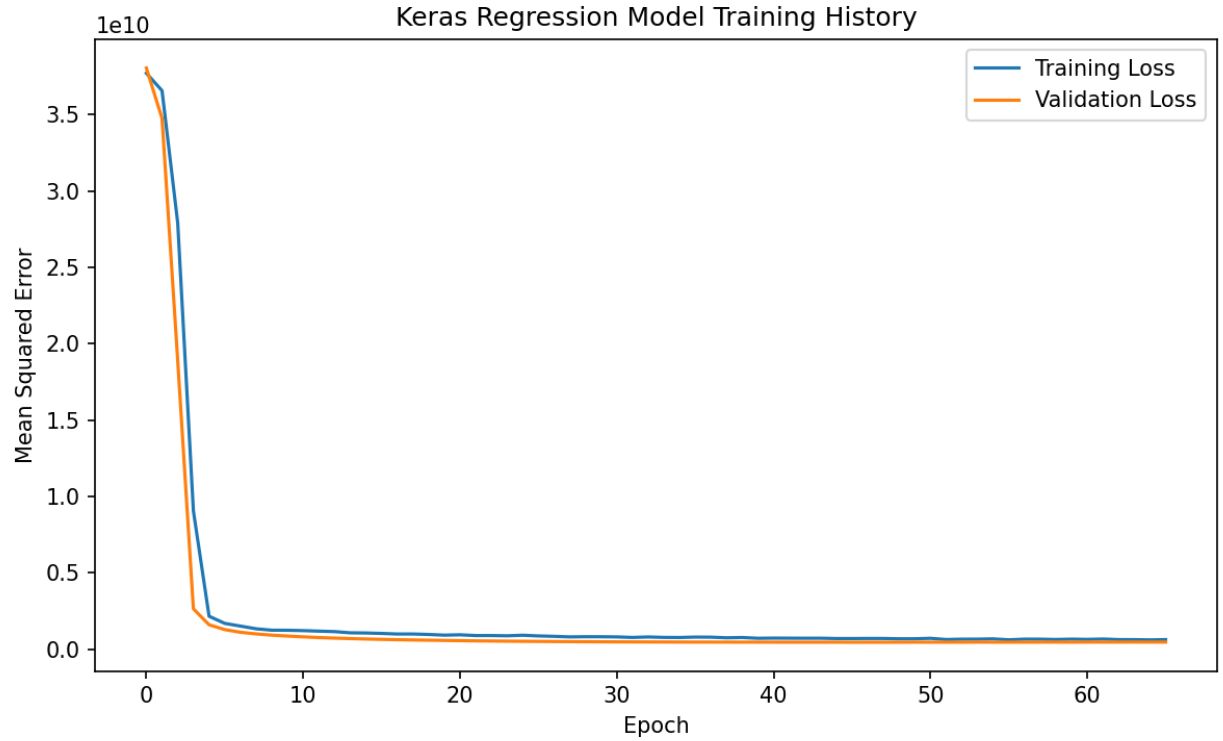


Figure 3: Keras regression model training and Validation loss

Conclusion

This project selected regression as the main machine learning topic and applied it to house price prediction. The goal was to predict residential sale prices using housing characteristics from the Ames Housing dataset. The methodology used data

preprocessing, missing-value handling, feature scaling, and a Keras neural-network regression model. The main advantage of the regression approach is that it provides a numeric estimate of price based on many property features. This makes it useful for real estate valuation and pricing support. At the same time, prediction quality depends on the dataset, preprocessing, and how well the training data represents the homes being predicted.

References

Zhong, X. (n.d.). Regression [Lecture slides]. Department of Computer Science, University of Nebraska Omaha

Fannie Mae. (n.d.). Property valuation. Fannie Mae Single-Family.

Kaggle. (n.d.). House prices: Advanced regression techniques. Kaggle.

Scikit-learn Developers. (n.d.). Scikit-learn: Machine learning in Python. Scikit-learn.

TensorFlow. (2024a). Basic regression: Predict fuel efficiency. TensorFlow.