Machine Learning Models for Regression and Classification Datasets

Madison Poore, Spring 2025

Introduction

This paper will explore fitting standard and advanced Machine Learning models for two datasets – one regression dataset (housing data) and one classification dataset (loan data). For both datasets, this paper provides three fit models and an analysis on their performance. Please reference the associated Jupyter Notebook File for the script used to generate all metrics and figures.

Introduction – Housing Regression

The housing dataset focuses on predicting the median house value based on several different features. The features in question: longitude, latitude, median housing age, total rooms, total number of bedrooms, population, number of households, median income, and proximity to the ocean.

Regression Dataset – Linear Model

The linear model fit for the housing (regression) dataset is Polynomial Features. This training process consisted of splitting the data into 25% testing and 75% training sets. Subsequently, a 2nd degree polynomial features model was trained via Sci-Kit Learn Polynomial Features object. Finally, based on the training *and* testing data, we computed the Mean Squared Error and R^2.

The training set MSE is 3.79e9, and the testing set MSE is 4.42e9. It is again important to note that these metrics are computed in the context of housing prices. The relative error between the training and testing set is less than 20%, which indicates that the Polynomial Features Model generalizes well to unseen (testing) data.

The training set R^2 is 0.715, and the testing set R^2 is 0.666. We desire an R^2 that is close to 1, and both training and testing metrics are reasonably close to 1. Since 0.666 is so close to 0.715, there is not much risk of overfitting.

Regression Linear Model SHAP Evaluation

SHAP (SHapley Additive exPlanations) allow us to understand the output of this linear model by quantifying the contribution of each feature present. That is, the variables that have a larger (in magnitude) SHAP value will be more impactful to the output of the model. In the context of this problem, the proximity to the bay is not very impactful to the model while the total number of rooms is very impactful.

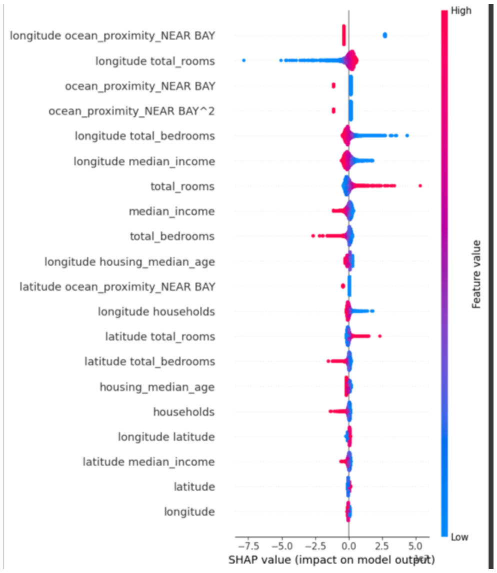

Figure 1: Polynomial Features Model SHAP Evaluation

Regression Dataset Model Advanced Machine Learning Models

I. Decision Trees

The first advanced machine learning model fit for the regression dataset is Decision Trees – chosen for its explainability. Decision Trees are used for both Regression *and* Classification datasets using the Sci-Kit Learn Decision Tree Regressor. The structure of a tree is determined during training (and hence by the data itself). The results of the Decision Tree training are evaluated using MSE and R^2.

The Decision Tree MSE is 3.71e9 and the Decision Tree R^2 is 0.719. The most important features according to the Decision Tree Model is median income.

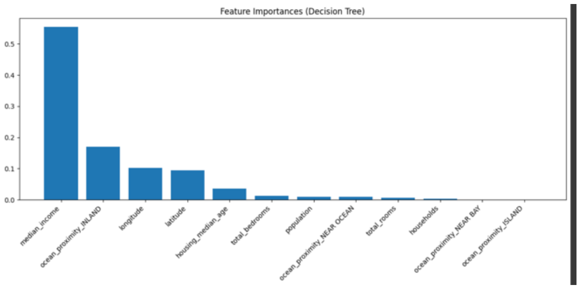

Figure 2: Decision Tree Feature Importance

II. Neural Networks

The final model fit for the regression (housing) dataset is a Neural Network within Sci-Kit Learn. While Neural Networks are typically known as a 'black box', we can still use the SHAP evaluation to increase explainability of the model and get direction for tuning parameters.

The MSE generated by the Neural Network is 4.14e9 and the R^2 is 0.686. Again, this does raise concern in the context of generalizing the model or in overfitting.

The SHAP plot tells us that the most important features to the model are longitude, latitude, population, median income and households. It is important to recognize how different models place value on different variables. For example, in the linear model, there was a high weight placed on the total rooms variable. However, with the decision tree model, there is less of a focus on total rooms and much more of a focus on median income.

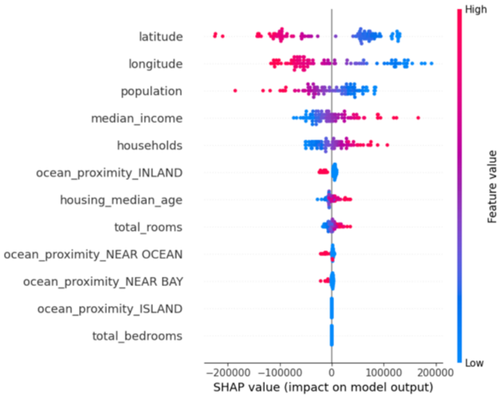

Figure 3: Neural Network SHAP values

Regression Dataset Conclusions

Having fit three distinct models (Polynomial Features, Decision Tree, and Neural Network), we can compare MSE and R^2 to determine which model has the best performance.

	MSE	R^2
Polynomial Features	3.79e9	0.715
Decision Trees	3.71e9	0.72
Neural Networks	4.14e9	0.69

The Decision Tree Model demonstrates the best performance with the lowest MSE and the highest R^2. Further tuning of these models could include removing features that do not have significant impact on future decisions and subsequently iterating on this process. This could mean excluding variables ocean proximity and households to create a simpler model overall.

Introduction – Loan Classification

The loan dataset focuses on predicting loan approval status (approved or rejected) by looking at 45,000 observations. There are several predictors present in this dataset including age, gender, highest level of education, income, ownership, years of employment experience, amount of loan, purpose of the loan, interest rate, percentage of income, length of credit history, credit score, indicator of previous loan defaults, and loan status (which is the target variable).

Again, we fit three machine learning models: Logistic Regression, Decision Trees, and Neural Networks. All three of these models are appropriate in the context of a classification model.

Classification Dataset – Linear Model

The linear model appropriate for a classification dataset is a Logistic Regression Model. Once again, we split the data into 25% testing and 75% training sets. After the model fit, we can cite the classification report to understand how accurate our model is. Subsequently, we can compute SHAP values to understand how to iterate beneficially on this Logistic Regression model.

Starting with the analysis of the classification report, we can draw some conclusions about the overall performance of our current Linear Model.

	Precision	Recall	f1-score	Support
Class 0	0.86	0.94	0.90	8730
Class 1	0.70	0.47	0.57	2520
Accuracy			0.84	11250
Macro average	0.78	0.71	0.73	11250
Weighted average	0.83	0.84	0.83	11250

Starting with the first row, Class 0, we can see a precision of 0.86, which means when our model predicts approvals, it's right 86% of the time. The recall rate (0.94) means the model is catching 94% of all true approvals. An f1-score of 0.90 indicates a strong performance for the model in terms of rejecting loan applicants. These metrics mean the Logistic Regression model is good at knowing when to approve a loan applicant.

With respect to class 1 (rejections), we have a precision of 0.70, meaning 70% of predicted rejections are correct. A recall rate of 0.47 indicates the model only catches 47% of actual rejections. An f1-score of 0.57 indicates there is some room for improvement in predicting rejections for our model. In short, we are approving too many people for loans.

Combining these two factors, we can see that our model is biased towards approving applicants on loan applications. That is, the model is better at predicting acceptance than rejection and is more likely to accept someone with a loan that should not have been approved. We will take note of this and see if our Advanced Machine Learning Models will provide better metrics in the following sections.

Classification Linear Model SHAP Evaluation

Some next steps for our model selection process are computing the SHAP values for our Logistic Regression Model to understand which features contribute to our prediction the most. Recall the magnitude of our SHAP values indicate how influential the model views a feature to be when predicting the output.

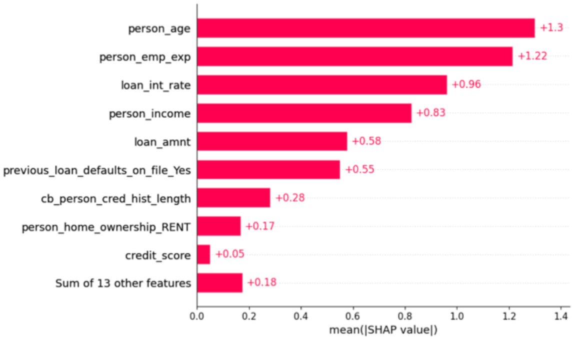

Figure 4: SHAP Values Logistic Regression

This bar plot demonstrates the SHAP values for each feature included in the model. From the bar plot, we can see a person's age is the most influential feature for this model, while credit score and other not listed features are not very impactful. However, in our performance analysis of this Logistic Regression model, we showed there is room for improvement when predicting the approval status. Because of this, we look to more advanced Machine Learning Model Methods in hopes of eliminating this bias.

Classification Dataset Advanced Machine Learning Models

Mirroring what was done for the Regression dataset to start, we will implement two separate advanced machine learning models (Decision Trees and Neural Networks) in the context of loan status.

I. Decision Trees

The first advanced machine learning model presented for the classification (loan) dataset is decision trees, again chosen for explainability. This analysis leans on Sci-Kit Learn's Decision Tree Classifier object rather than the Decision Tree Regressor as seen in the regression dataset.

The results of running the Decision Tree algorithm for this dataset are expressed in the following classification report.

	Precision	Recall	f1-score	support
Class 0	0.92	0.96	0.94	8730
Class 1	0.85	0.72	0.78	2520
Accuracy			0.91	11250
Macro average	0.89	0.84	0.86	11250
Weighted average	0.91	0.91	0.91	11250

Starting with the first row, Class 0, we can see a precision of 0.92 which means when our model predicts approval, it's right 92% of the time. The recall rate (0.96) means the model is catching 96% of all true approvals. An f1-score of 0.94 indicates a strong performance for the model in terms of approving loan applicants. It is important to note that all these metrics are an improvement from our Logistic Regression Model Classification Report and the bias between classifications has decreased.

With respect to class 1, we have a precision of 0.85, meaning 85% of predicted rejections are correct. A recall rate of 0.72 indicates the model catches 72% of actual rejections. An f1-score of 0.78 indicates a moderately strong performance of our model. Once again, these metrics are an improvement from our Logistic Regression Model. That is, we should favor the use of the Decision Tree model over Logistic Regression Model.

II. Neural Networks

A neural network is the final advanced Machine Learning Model we will fit for the loan dataset. Using Sci-Kit Learn's Neural Network MLP Classifier, we can train the model on 75% of the loan data. Subsequently, we can use the classification report to compare the performance of this model to the performance of previous models (Logistic Regression and Decision Trees).

	Precision	Recall	f1-score	support
Class 0	0.85	0.95	0.90	8730
Class 1	0.70	0.42	0.53	2520
Accuracy			0.83	11250
Macro average	0.78	0.68	0.71	11250
Weighted average	0.82	0.83	0.81	11250

Like our Logistic Regression Model, the Neural Network seems to be doing well at approving people from the loan. This is shown from a high precision and recall rate for Class 0 in the chart above. However, the low recall for Class 1, which is rejections, tells us that the model misses many people who should have been rejected. The precision of 0.70 for Class 1 (rejection) is the same as our Logistic regression, but the recall and f1-score are the worst for the neural network.

Classification Data Conclusions

Choosing between the three models (Logistic Regression, Decision Trees, Neural Networks), we can choose the model that has the best (highest) precision, recall and f1-score between approvals and rejections (Class 0 and Class 1). The “best” model for the loan data is the Decision Trees Model. For Class 0, the Decision Tree’s precision is 0.92, recall is 0.96 and f1-score is 0.94. For Class 1, the Decision Tree’s precision is 0.85, recall is 0.72 and f1-score is 0.85. All metrics shown for Decision Trees are higher than the other two models (Logistic Regression and Neural Networks). The decision tree model has the highest “balance” between accurate rejections and approvals and therefore we choose this model to make predictions for us in the future.