Grace McPadden

Professor Fei Miao

CSE 4820

3 May 2024

Machine Learning Applications in Health and Consumer Cost Prediction

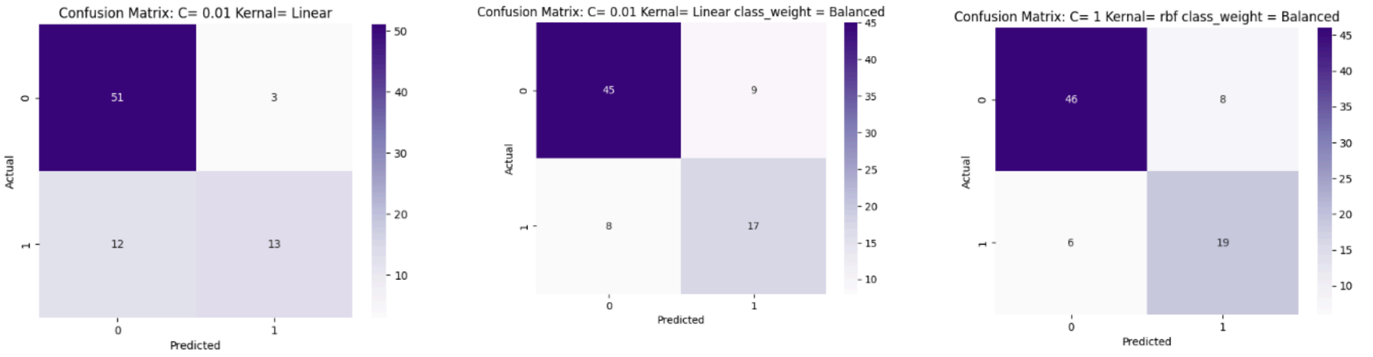
This analysis explores the use of supervised machine learning models on two datasets—one concerning diabetes diagnosis, and the other predicting customer acquisition costs. The objective of the analysis is to evaluate and compare the performance of multiple machine learning models, including Support Vector Machines (SVM), Logistic Regression, K-Nearest Neighbors (KNN), Linear Regression, and a Neural Network using Multilayer Perceptrons (MLP). Each dataset required appropriate preprocessing and model tuning, and the results were evaluated using accuracy, R^2 score, precision, recall, and visualizations. This report summarizes the methods, performance, and takeaways from applying machine learning models to classification and regression problems.

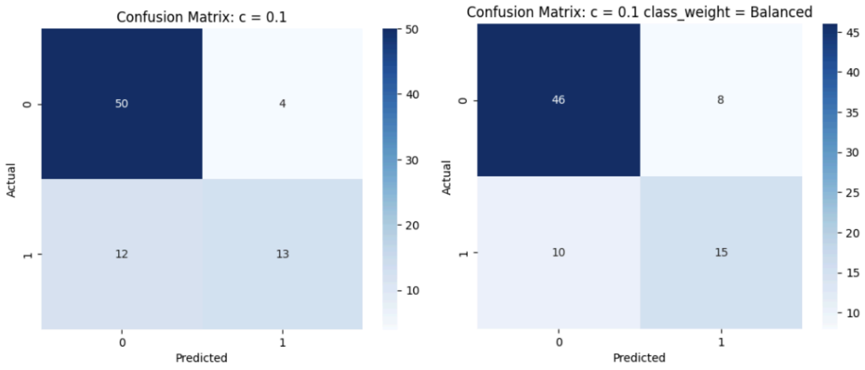
The first dataset used in this analysis is from the National Institute of Diabetes and Digestive and Kidney Diseases. It contains various diagnostic measurements and a binary classification of whether the patient was diagnosed with diabetes. It is important to note that this specific data set only contained female patients over 21 of Pima Indian heritage, so the results of this analysis are only generalizable to this subgroup. The objective of the analysis was to use supervised learning to find the best possible classification model for predicting whether a patient has diabetes.

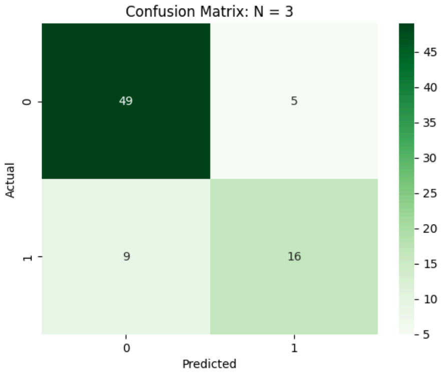
This data set contained several parameters along with a binary classification for whether a patient was diagnosed with diabetes. The parameters measured were the number of pregnancies, glucose level in the blood, blood pressure, skin thickness, insulin, BMI, diabetes pedigree score, and age. There are 768 patients in this data set prior to preprocessing. Unfortunately, some of the columns had missing entries. For instance, there were 35 patients in this data set who had a blood pressure of 0, which is not possible. To fix this, patients who had a 0 in insulin, blood pressure, glucose, skin-thickness, BMI, or age were removed. This left 392 patients after filtering. The second step of preprocessing was to normalize the data so each parameter had a mean of 0 and a standard deviation of 1.

The first model that was applied to the data set was an SVM. The parameters that were adjusted were the kernel and the C (regularization parameter). After running the model through multiple iterations with different parameters, the highest accuracy for test classification was a 0.81 with a linear kernel and a c of 0.01. However, upon looking at the confusion matrix it was apparent that although the model had high precision (0.81), it had very low recall (0.52). So while it had a low amount of false negatives, the number of false positives was very high. One possible reason for this is that the data set is imbalanced and 66% of the samples are a negative diagnosis. Despite high accuracy, these parameters may have caused underfitting due to the simple kernel and high regularization. To try to raise the recall, the parameter class_weight was set to balanced. This automatically assigns class weights inversely proportional to class frequencies. After running the model again with this parameter with the linear kernel and c = 0.01, the recall was raised from 0.52 to 0.68. For a second trial varying the parameters with the class_weight set to balanced, the model with the highest accuracy had a C = 1 and a kernel= rbf. For these parameters, the overall accuracy was 0.82 with a precision of 0.88 and a recall of 0.85.

The second model applied to this data set was logistic regression. Different C values were iterated through to find which had the highest accuracy. With the class_weight set to its default value, the accuracy was 0.81. This model had a high precision (0.76) but low recall (0.52). With the class weight set to balanced, the overall accuracy was 0.78 and the recall was slightly higher with the balanced class weight at 0.60. However, the precision was lower at just 0.65. The third model I tested was K-nearest neighbors. I tested a variety of k parameters and found that the highest accuracy was obtained with k = 3. The accuracy for this model was 0.83 with a precision of 0.76 and a recall of 0.64.







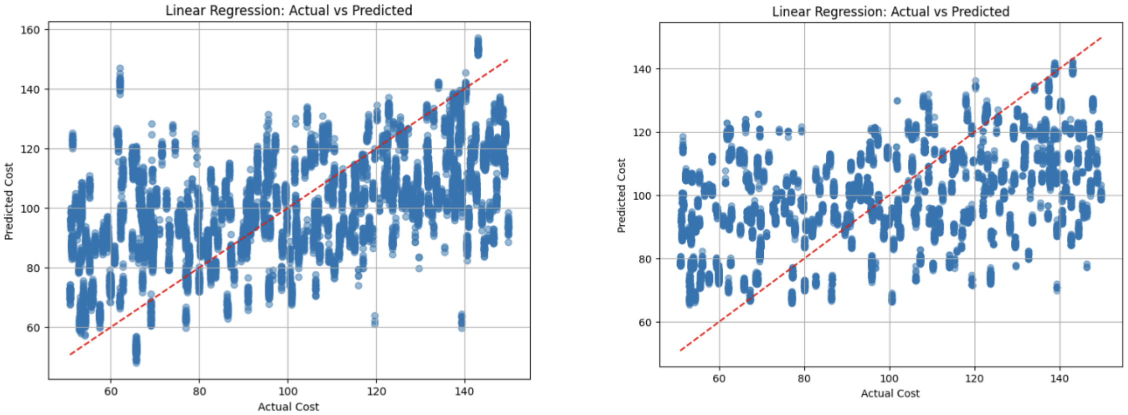
First Figure (Purple): Confusion Matrices for SVM, Second Figure (Blue): Confusion Matrices for Logistic Regression, Third Figure (Green): Confusion Matrices for KNN

Overall, SVM appears to be the best model for classification on this dataset. Through experimentation, the optimal parameters were found to be C = 1, kernel = rbf and class_weight = balanced, with an overall accuracy of 0.82, precision of 0.88, and recall of 0.85.

The second data set used in this analysis contains various metrics on a sale with the cost of acquiring that customer. In total, there were 36 variables measured for each customer including the type of food purchased, store sales, customer income, and marital status. The objective of this problem was to use supervised learning to solve the regression problem to predict the cost of acquiring a customer from these variables. In total, there are 60,429 customers in this data set.

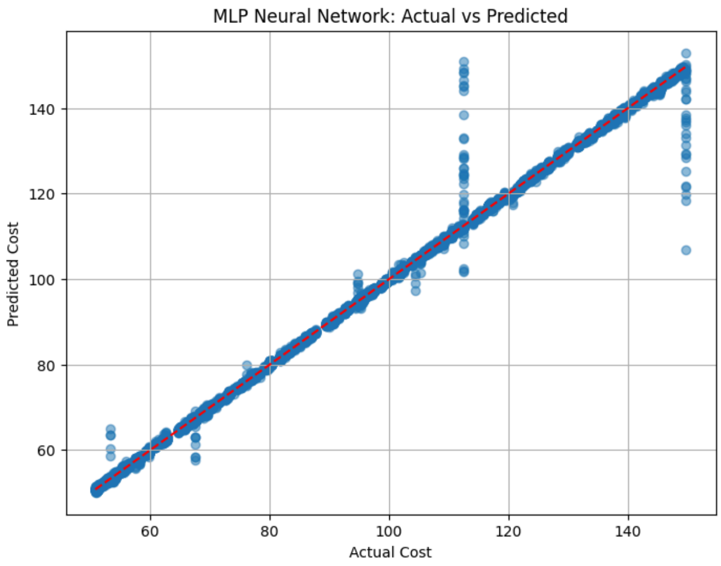
Many of the categories contain non-numerical labels, for instance, marital status or promotion name. As part of preprocessing, these labels were converted into numerical values. The two methods of doing this conversion are using the OneHotEncoder or OrdinalEncoder functions of the sklearn module. The OrdinalEncoder method assigns arbitrary values to each category (ex. A= 1, B = 2, C=3), which the model may interpret as order. OneHotEncoder avoids this by converting categorical columns into a binary matrix. The OneHotEncoder was used for categories without order, such as gender or promotion_name. The OrdinalEncoder was used for categories with ordering, such as income range or level of education.

The first model that was applied to this data set was a linear regression. After training the model using all of the features, the train R^2 was 0.356 and the test R^2 was 0.352. This indicates that the model is not overfitting, but also that the model explains just 35.6% of the variance. For the second trial with linear regression, the features used to train the model were reduced from 36 to 11 features. The categories that were included were the promotion name, sales country, marital status, occupation, home ownership, education, member card, yearly income, total children, and amount of cars owned. This reduction of the number of features reduced the test R^2 to just 0.269 -- which shows that less obviously relevant features (such as package recyclable or store square feet) still have predictive power on customer acquisition costs.



First Figure: Actual Versus Predicted Cost Using all features, Second Figure: Actual verus predicted cost using reduced features.

The second model applied to this data set was a neural network. The network used was a Multilayer Perceptron. Two hidden layers were used for this model, the first had 64 neurons and the second had 32 neurons. The reLU activation function was used with the adam optimizer, and there were a total of 500 iterations. This model had a much higher predictive power than the linear regression with a train R^2 of 0.999 and a test R^2 of 0.998. This indicates that over 99% of the variation in the data is explained by the model. The test and train predictive power are similar, which indicates that the model is not overfitting. The neural network using the MLP model had a much higher predictive power than the linear regression. This is likely due to the linear regression model not being complex enough to capture the variation in the dataset.



First Figure: Actual Versus Predicted Cost Using Multilayer Perceptron Neural Network.

In summary, this analysis demonstrates the importance of model selection and preprocessing in supervised learning tasks. For the classification problem, the Support Vector Machine with an RBF kernel and class balancing outperformed Logistic Regression and KNN in both precision and recall. For the regression task, the neural network model (MLP) outperformed linear regression with its ability to capture complex relationships in the data. These findings highlight the power of the model section as well as feature encoding, model tuning, and evaluation metrics in building effective predictive models.

Code:

Predicting Diabetes Diagnosis

https://colab.research.google.com/drive/15lk3YsAZSWvomyYiJS0LmQ8gPtejh3ca?usp=sharing

Predicting Customer Acquisition Costs

https://colab.research.google.com/drive/1RXfqW4xRy6D7m8zrzYYRVTDQRguIG1kL?usp=sharing

Datasets:

https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset

https://www.kaggle.com/datasets/ramjasmaurya/medias-cost-prediction-in-foodmart