DIAML: Assignment 5

by Anhtony Rizkallah

Table of contents

1 Statistical Learning . . . 3

1.1 Rule-Based Approach to Decision-Making . . . 3
1.2 Over Fitting . . . 3
1.3 Methods of Avoiding Over Fitting . . . 4
1.4 Model Performance Evaluation Metrics . . . 4
1.5 Benchmarks . . . 4

2 Machine Learning . . . 4

2.1 What is Machine Learning? . . . 4
2.2 Machine Learning Techniques . . . 5
2.3 Classification vs Regression . . . 5
2.4 Supervised vs Unsupervised Learning . . . 5
2.5 Machine Learning Applications . . . 5

3 Diabetes Data . . . 6

3.1 Explanatory Variables Correlation Matrix and Heatmap . . . 6
3.2 Collinearity . . . 7
3.3 Linear Model . . . 7
3.3.1 Steps . . . 7
3.3.2 Results . . . 8
3.4 Forward vs Backward Selection . . . 8
3.5 Stepwise . . . 8
3.5.1 Methodology (p-value entry and exit evaluation) . . . 9
3.5.2 Methodology (BIC -or any other metric) . . . 9
3.6 Insights . . . 10

4 Analyzing the Titanic Dataset . . . 11

4.1 Logistic vs Linear Regression . . . 11
4.2 Probabilty of Survival . . . 11
4.2.1 Steps . . . 11
4.2.2 Results . . . 11
4.3 Probabilty of Survival by Age, Gender, and Passenger Class . . . 11
4.3.1 Steps . . . 11
4.3.2 Results . . . 12
4.4 Logistic Regression Model . . . 12
4.4.1 Steps . . . 12
4.4.2 Results . . . 12
4.5 Model Performance . . . 13
4.5.1 Steps . . . 13
4.5.2 Results . . . 13

Python Packages

Pandas

Sklearn.metrics

Sklearn.model_selection

Sklearn.linear_model

Stepwise_regression

Statsmodels

Scipy

Matplotlib

Seaborn

1 Statistical Learning

1.1 Rule-Based Approach to Decision-Making

1. Defining Objectives: Developing a purpose and expected outcome from the decision-making process
2. Identifying Key Variables: Determining the factors that influence the decision
3. Setting Rules: Creating rules based on conditions for each key variable
4. Applying and Testing the Rules: Testing that the rules are yielding expected outcomes
5. Revisit Framework: If the test did not yield expected outcomes, revisit variables and rules

For example, in a loan-approval system, the objective is to determine whether an applicant qualifies for a loan (Step 1). So, for a loan approval, assume that the identifiying key variables include income, credit core, and debt-to-income ratio (Step 2). After the variables have been identified, the evaluation rules are (Step 3): approve if

1. Income $>$ 80,000 USD, Credit Score $>$ 750, Debt-to-Income ratio $<$ 3
2. 50,000 USD $<$ Income $<$ 80,000 USD, Credit Score $>$ 750, Debt-to-Income ratio $<$ 1.5

An applicant with an income of 75,000 USD, credit score of 760, and debt-to-income ratio is 1.5 gets approved for the loan after meeting rule number 2 (Step 4). However, after getting multiple applications, 90% of the applicants did not meet the criteria and the loaner decided to revisit and loosen the rules to loan more money out to increase profit (Step 5).

Domain Knowledge

Domain knowledge is essential in establishing rules -or at least effective ones- because it brings relevant insight to factor in/out variables and what the various impacts would be. For example in Step 5, for the loaner to revisit the rules, they must understand the credit risks associated with the different variations of the loan-approval system. If someone with no domain knowledge attempts to build the rules, they won't know what the explanatory variables are.

1.2 Over Fitting

Over fitting is when a statistical model learns the noise and details from the training data rather than learning the general patterns which leads to memorization. This causes the model to follow every point in the training dataset with almost a perfect match which eventually begins to fit the noise. Then, when the model is applied to the test dataset, it is very likely to fail in accurately predicting values.

With a small dataset of ten data points, adopting a simple model with one parameter is definitely preferred over a complex model with 10 parameters because it is more likely to learn the general trend in the data rather than memorize.

1.3 Methods of Avoiding Over Fitting

Cross Validation: This procedure divides the dataset into training and testing sets multiple times, and trains the model on each subset. It avoids over-fitting by ensuring that the model was validated across multiple datasets and reduces the risk of over-fitting.

Regularization: This process adds a penalty to the model based on its complexity, meaning, that the simpler the model (less variables), the less penalty it will have, and the more general it will be.

1.4 Model Performance Evaluation Metrics

- Mean Absolute Error: $\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} |\hat{y_i} - y_i|$, where N is the number of data points, $\hat{y_i}$ is the predicted value, y_i is the actual value
- $\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \times 100$

Applications

- In a *regression model* that predicts house prices, MAE should be used to evaluate the performance because it can evalute the prediction in dollars.
- In a *classifcation model* that diagnoses patients for a disease, accuracy measures how many times the model correctly predicted the disease.

1.5 Benchmarks

Benchmarks are useful in evaluating model performance objectively, and compare how well the model performs relative to others- accuracy, precision, explanatory power, etc. Moreover, benchmarking is useful in identifying model strengths and weaknesses, meaning that the model can underperform on certain datasets and overperform on others, and using benchmarks can help find them and adjust the model. Additionally, decision makers that rely on machine-learning models have certain KPIs to achieve, and benchmarks are helpful in assessing how the model is performing to achieve specific outcomes, giving insights to decision makers about how the model is operating so they understand what their decision involves. Examples in different contexts:

1. In a business setting, benchmarks include specific goals that the model should achieve. For example, achieving advertisement conversions of 40% from website traffic. The model can then be evaluated based on the conversion rate, ensuring compliance with the business goals.
2. Statistically, there are publicly available datasets, such as IMDB sentiment analysis[1], which come with already established benchmarks. Then, these datasets can be used to measure model performance and compare them directly with other ones and improve the current model accordingly.

2 Machine Learning

2.1 What is Machine Learning?

Machine Learning is a subset of Artificial Intelligence that enables computers to learn from data and predict future outomces. Machine Learning algorithms learn trends in historical data, and apply their understanding of those trends to predict the future. With the rise of computers, what seemed to be lengthy and time-consuming processes that were previously accomplished by a labor force can now be achieved through computers. This revolution carried with it a lot of benefits, such as efficiency, accuracy, and technology, along with Machine Learning applications. In 1950[2], Alan Turing developed a test that can measure wether a machine can demonstrate intelligence. He discovered that a machine can actually learn, and when communicated with it, it is indistinguishable from a human. Later in 1952, Arthur Samuel of IBM wrote the first game-playing program, ELIZA to achieve sufficient skill to challenge a Checkers World Champion. In 1957, Frank Rosenblatt, a professor at Cornell University, invented perceptron, a simple linear classifier. In 1990, computer science and statistics combined to provide data-driven approaches to

1. https://paperswithcode.com/sota/sentiment-analysis-on-imdb
2. 18-785: Data, Inference, and Applied Machine Learning Week 7 Slides

machine learning. In 1997, one key milestone is the famous Deep Blue vs Garry Kasparov match, where it was the first time that a chess grandmaster loses to a computer. Deep Blue relied on brute-force search algorithms and handcrafted evaluation functions, rather than learning from data. Although it did not use machine learning, Deep Blue represented an achievement in AI, proving that computers can outperform experts in complex tasks. In 2010, Big Data made a breakthrough with exponential growith in the volume, velocity, and variety of data available for analysis and research. In 2014, infrastructure, protocols, and standards for providing open access to data via APIs gained a lot of traction. Today, Machine Learning has gained great popularity, and in terms of chess, it was used to build models such as Google AlphaZero, which learned chess strategies by playing games against itself using reinforcement learning. ML's ability to uncover patterns, make predictions, automate complex tasks in real-time applications, and applicability to various fields gave rise to its popularity. Almost all fields that require forecasting, predictions, and interdisciplinary problem-solving use ML models because it can either uncover uncharted territory or reinforce current knowledge by making more accurate predictions by using statistical methods.

2.2 Machine Learning Techniques

- **Classification (Supervised Learning)**: Assigns categorical labels to datapoints based on different input features.
- **Regression (Supervised Learning)**: Predicts continuous values by finding patterns and/or trends in data.
- **Clustering (Unsupervised Learning)**: Grouping of similar data points together without predefined labels. In other words, it can categorize data together without any input features.

2.3 Classification vs Regression

	Classification	**Regression**
Types of Output	Categorical (e.g. Yes/No)	Numerical (e.g. 100, 200, etc.)
Evaluation Metrics	Accuracy, Precision, F-score	MAE, MSE, R^2, MAPE
Objective	Identify and predict categories	Model and predict continuous values

Table 1. Classification and Regression Differences

2.4 Supervised vs Unsupervised Learning

	Supervised Learning	**UnSupervised Learning**
Input Data	Data has labels	Data does not have labels
Data Usage	x features and Y variables: Y = f(x)	Finds patterns in the x features w/out Y
Objective	Predict outcomes based on training data	Find hidden patterns in data

Table 2. Supervised and Unsupervised Learning Differences

2.5 Machine Learning Applications

Application	**Technique**	**Type of Learning**	**Description**
Predicting wind speed at a wind farm	Regression	Supervised Learning	Predicts wind speeds using historical weather data and environmental variables
Medical diagnosis of heart attack	Classification	Supervised Learning	Classifies patients as high or low risk for heart attack based on symptoms, history, and BMI
Customer segmentation in Marketing	Clustering	Unsupervised Learning	Groups customers based on purchasing behaviors and/or demographics to target them with relevant ads and/or specific search results

Table 3. Supervised and Unsupervised Learning Differences

3 Diabetes Data

3.1 Explanatory Variables Correlation Matrix and Heatmap

	AGE	SEX	BMI	BP	S1	S2	S3	S4	S5	S6
AGE	1.0000	0.1737	0.1851	0.3354	0.2601	0.2192	-0.0752	0.2038	0.2708	0.3017
SEX	0.1737	1.0000	0.0882	0.2410	0.0353	0.1426	-0.3791	0.3321	0.1499	0.2081
BMI	0.1851	0.0882	1.0000	0.3954	0.2498	0.2612	-0.3668	0.4138	0.4462	0.3887
BP	0.3354	0.2410	0.3954	1.0000	0.2425	0.1855	-0.1788	0.2577	0.3935	0.3904
S1	0.2601	0.0353	0.2498	0.2425	1.0000	0.8967	0.0515	0.5422	0.5155	0.3257
S2	0.2192	0.1426	0.2612	0.1855	0.8967	1.0000	-0.1965	0.6598	0.3184	0.2906
S3	-0.0752	-0.3791	-0.3668	-0.1788	0.0515	-0.1965	1.0000	-0.7385	-0.3986	-0.2737
S4	0.2038	0.3321	0.4138	0.2577	0.5422	0.6598	-0.7385	1.0000	0.6179	0.4172
S5	0.2708	0.1499	0.4462	0.3935	0.5155	0.3184	-0.3986	0.6179	1.0000	0.4647
S6	0.3017	0.2081	0.3887	0.3904	0.3257	0.2906	-0.2737	0.4172	0.4647	1.0000

Table 4. Explantory Variables correlation matrix rounded to the nearest ten thousandth

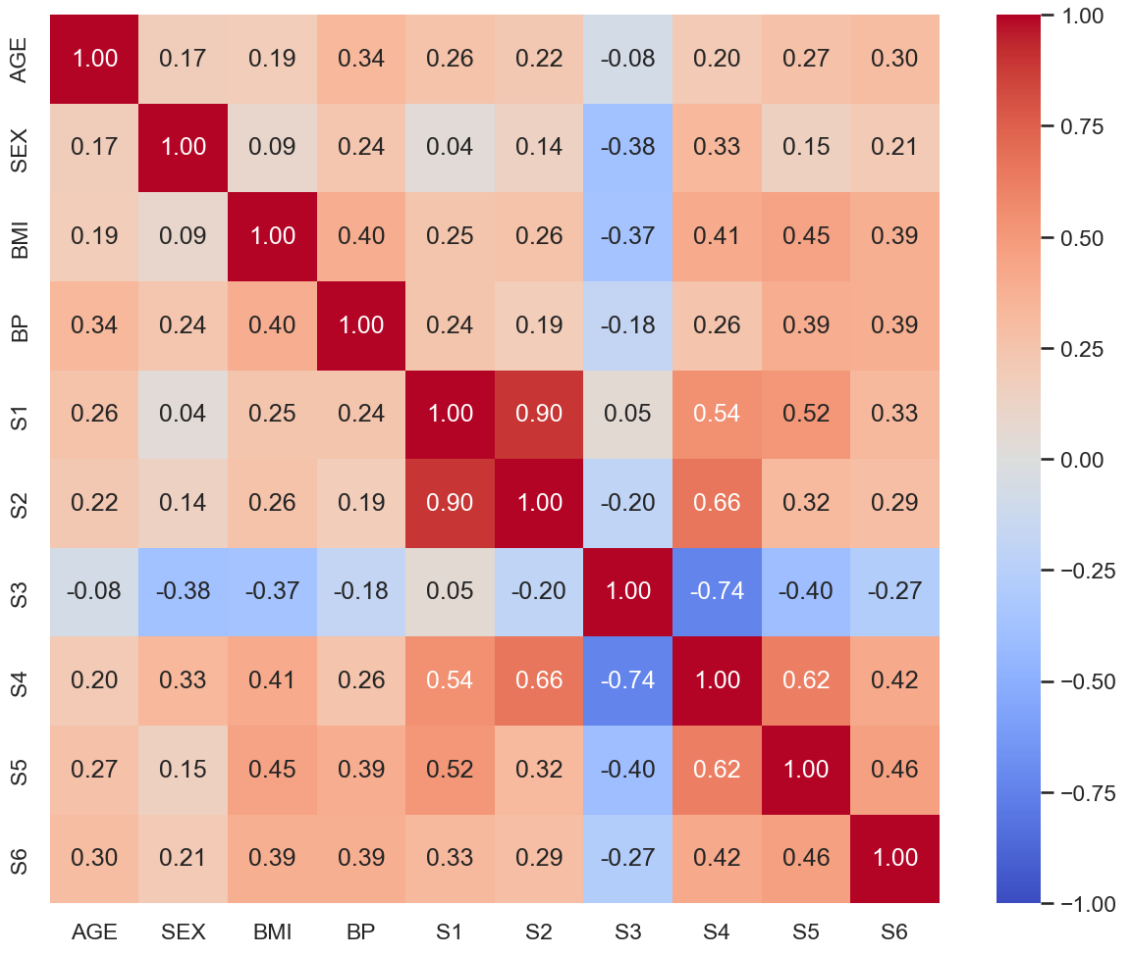

Figure 1. Explanatory Variables correlations heatmap rounded to the nearest hundredth

To desribe the correlation between the different explanatory variables, classifying them as high, moderate, and low to understand the relevance of their correlations is relevant.

- High Correlations (x>0.7): S1 and S2 have a high positive correlation which means that the serum measurements are strongly related to each other. This is possibly because the first

two measurements were done within a short timeframe, so the correlation might be high due to the serum levels being close to each other.

- Moderate Correlations (0.5<x<0.7): The serum levels S1, S2, S3, S4, and S5 are moderately correlated with each other. This is also due to the measurements being done within the timeframes, so the fact that there exists a moderate correlation makes sense.

- Low Correlations (x<0.5): The remaining variables are minimaly correlated with each other, showing low or near-zero correlation values. This is a good sign, showcasing independence of variables from each other.

3.2 Collinearity

Collinearity is the correlation that exists between independent variables (predictors). This means that the variables have a relationship with each other. In other words, multicollinearity is the term used to describe this dependence among multiple intercorrelated predictors. This is generally problematic because independent variables should be independent, similar to the linear algebra concept of linear dependence among elements in a matrix.

So, the effects of collinearity among predictors can have counterproductive effects on regression models. In principle, each predictor is associated with a $\widehat{B}_i$ value in a regression model. For example:

$$\hat{y} = \hat{B}_0 + \hat{B}_1 x_1 + \hat{B}_2 x_2 + \cdots \widehat{B}_i x_i + \epsilon$$

where $\widehat{B}_0$ is the intercept, $\widehat{B}_i$ is the estimated coefficient for the predictor x_i, and ϵ is the residual. If the variables are correlated, the estimated coefficients won't be distinct– each predictor should fit a separate piece of the dependent variable. As collinearity increases, it becomes more difficult to determine the effect of each variable on the estimated coefficients which undermines the clarity of the model.

3.3 Linear Model

3.3.1 Steps

First, to build the model, separate the independent variables (predictors: everything but Y for this model) from the dependent variable (Y). Then, fit a regression model where the dependent variable is regressed onto the independent variables.

Mean Squared Error (MSE) and R^2 (or Adjusted R^2) are the metrics that can assess accuracy and goodness-of-fit, respectively.

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N} (\hat{Y}_i - Y_i)^2$$

where N is the number of data points, $\hat{Y}_i$ is the predicted values, and Y_i is the actual values.

$$R^2 = 1 - \frac{\text{SS}_{\text{res}}}{\text{SS}_{\text{tot}}}$$

where $\text{SS}_{\text{res}} = \sum (y_i - \hat{y}_i)^2$ and $\text{SS}_{\text{tot}} = \sum (y_i - \bar{y})^2$. y_i is the actual value, $\hat{y}_i$ is the predicted value, and $\bar{y}$ is the average value of the actual values.

R^2_{adj} involves a penalty system to balance the increased complexity in the model (more predictors) with an improved R^2. In other words, adding a predictor that does not significantly increase R^2 (explanatory power), then the R^2_{adj} will decrease with every added predictor. Essentially, Adjusted R^2 is a more conservative metric that helps prevent overfitting by penalizing unnecessary predictors.

$$R^2_{\text{adj}} = 1 - \frac{(1-R^2)(N-1)}{N-p-1}$$

where p is the number of predictors.

Then after evaluating the model metrics, the p-values (statistical significance) of each explanatory variable needs to be verified whether it is significant (α<0.05) or not (α>0.05).

3.3.2 Results

Model	**MSE**	$\boldsymbol{R^2_{\text{adj}}}$
model1	2859.700	0.507

Table 5. MSE and Adjusted R^2 of model1 rounded to the nearest thousandth

	p-value
AGE	0.8670306
SEX	0.0001041671
BMI	4.296391×10^{-14}
BP	1.024278×10^{-6}
S1	0.05794761
S2	0.1603902
S3	0.6347233
S4	0.2734587
S5	1.555899×10^{-5}
S6	0.3059895

	p-value >0.05
AGE	0.8670306
S1	0.05794761
S2	0.1603902
S3	0.6347233
S4	0.2734587
S6	0.3059895

	p-value<0.05
SEX	0.0001041671
BMI	4.296391×10^{-14}
BP	1.024278×10^{-6}
S5	1.555899×10^{-5}

Table 6. Statistical significance of the explanatory variables for model1

The signficant variables are SEX, BMI, BP, and S5 because their p-values are less than 0.05. With only a handful of significant variables, collinearity is certainly an issue with model1. By the same token, the collinearity among the insignificant variables is very high which renders the model unstable. On the other hand, the significant variables have low collinearity with each other, which could make the model more stable, more interpretable, and better at generalization.

3.4 Forward vs Backward Selection

Both forward and backward selection methods are used to build a regression model. Forward selection starts with no predictors and adds one variable at a time. At each step, the variable that improves the model the most, minimizing regularization metrics (e.g. AIC, BIC) is added. Backward selection starts with all predictors and removes the least significant variable one at a time, also verifying if the model improves using the same regularization metrics.

3.5 Stepwise

Stepwise is a method that combines both forward and backward selection and eventually stops when the optimal model has been found[3]. The method that determines how predictors enter and exit the model are based on the metrics used to evaluate the performance. Keeping predictors based

3. https://online.stat.psu.edu/stat501/lesson/10/10.2

on their p-values being less than the threshold of entry (forward selection), and removing existing predictors if their p-values exceed the threshold of exit (backward selection). Another method of Stepwise uses BIC or AIC as the "improvement" metric to decide whether to add or remove a predictor. The following examples are two of the evaluation methods in stepwise.

3.5.1 Methodology (p-value entry and exit evaluation)

1. Choose the variable (x_1) with the lowest p-value after calculating the correlation coefficient of all variables as a preliminary step.

2. Regress the dependent variable on x_1 and x_2, x_1 and x_3 ... x_1 and x_p. The variables in the model that have a p-value higher than the exit alpha $\alpha_{\text{exit}} = 0.15$ (chosen for this problem) are removed (Backward step).

3. Assuming that x_1 and x_2 remained, regress the independent variable onto the model containing both variables and repeat step 2 to evaluate the p-values.

4. Re-iterate the process until no addition nor removal changes the p-values.

3.5.2 Methodology (BIC -or any other metric)

1. Similar to stepwise regression, but rather choose the variable that has the lowest BIC of the model (x_1)

2. Regress the dependent variable on x_1 and x_2, x_1 and x_3 ... x_1 and x_p. The model that has the lowest BIC among the other models is chosen.

3. Assuming that x_1 and x_2 remained, regress the independent variable onto the model containing both variables and repeat step 2 to evaluate the BIC.

4. Re-iterate the process until no addition nor removal changes the BIC values.

The formula for calculating BIC is

$$BIC = N \ln\left(\frac{\text{RSS}}{N}\right) + k \ln N$$

where BIC is Bayesian Information Criteria, N is the number of observations (data points), k is the number of parameters, and RSS is the residual sum of squares.

Both processes at the end will keep the significant variables with respect to each model's evaluative system. Additionally, one can use a combination of both p-values and/or performance metrics in the stepwise regression to develop their own variable-selection rules.

For the model, the methodology that was used is **forward selection** based on p-value. The procedure of a forward selection in this case is as follows:

1. Choose the variable (x_1) with the lowest p-value after fitting a regression model of the dependent variable onto the different variables distinctly, such that p-values must be less than the threshold of entry (<0.05).

2. Combine x_1 with x_2, x_1 with x_3, ..., x_1 with x_p, and regress. The variable with the lowest p-value is then chosen. Assume x_1 and x_2 are the selected predictors for the current model.

3. Perform Step 2 again until all the different iterations have been covered. The model will have all the different predictors that meet the p-value < 0.05 criterium.

Selected Predictors	BMI	S3	S5	SEX	BP	S4
P-values in final model	6.83×10^{-15}	5.14×10^{-5}	2.83×10^{-11}	1.90×10^{-4}	4.37×10^{-7}	0.301

Model	**MSE**	$\boldsymbol{R^2_{adj}}$
model2	2906.602	0.503

Table 7. Selected predictors using forward selection in model2 and model performance metrics

3.6 Insights

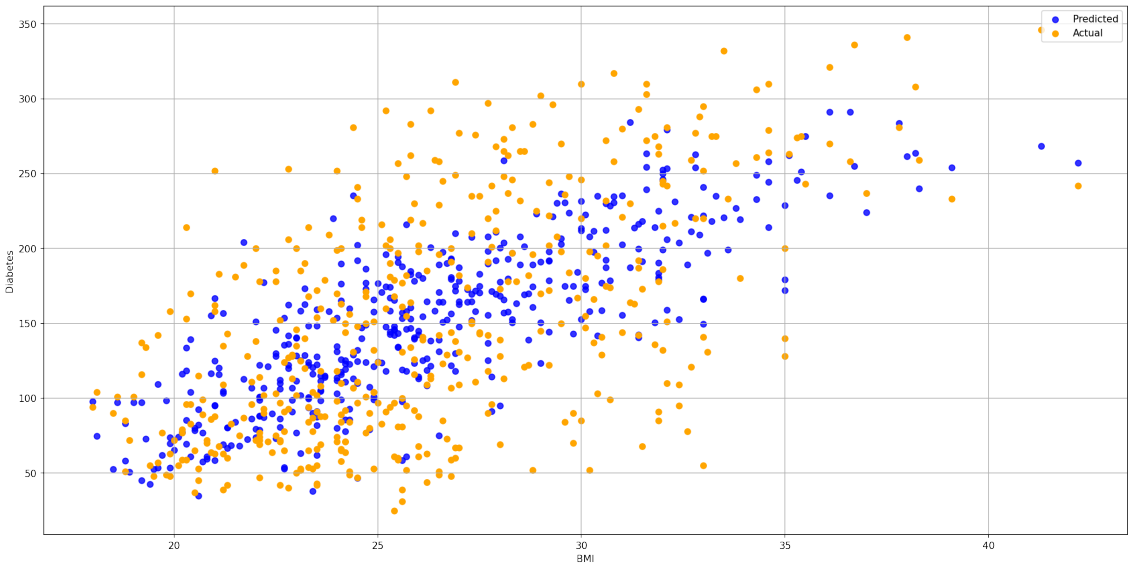

Figure 2. model1 scatterplot of actual and predicted values using all variables

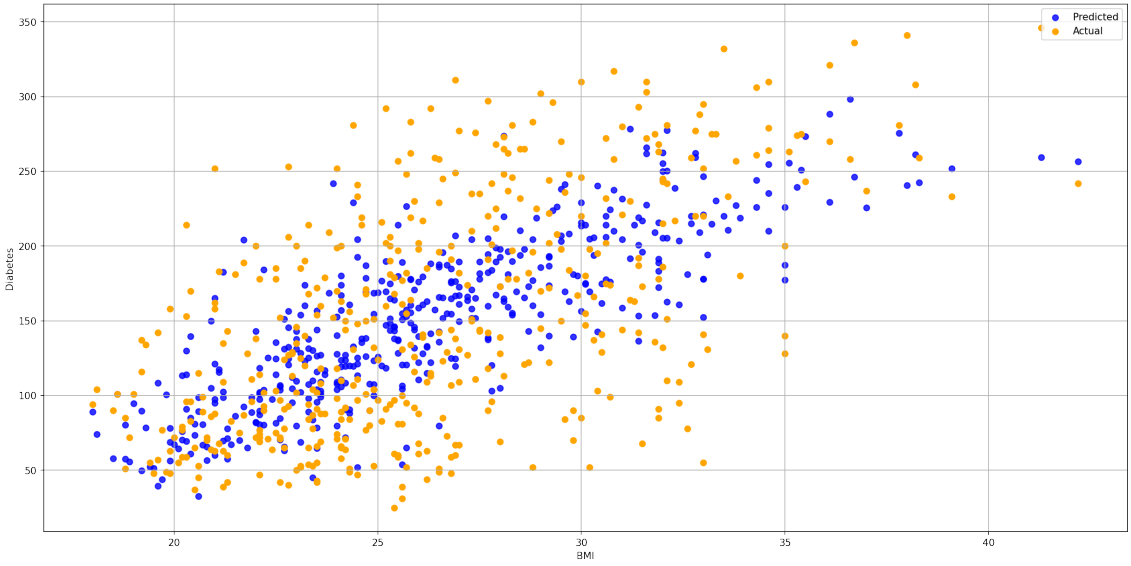

Figure 3. model2 scatterplot of actual and predicted values using forward selected variables

First, important to note is that the purpose of those plots is to show that both models have very close performances in terms how well they generalize. It can be seen that the model does not overfit, but rather generalizes- seen through the blue dots. Both models have an R^2 around 0.5 which means that they can explain about 50% of the variance in the data. Model1 has an MSE of 2859.7 which is lower than that of model2 at 2906.6. However, the tradeoff between both models is complexity, where model1's interpretability is ambigious and contains variables with no statistical significance, whereas model2 only contains 1 statistically insignifcant variable (S4: p-value of 0.301).

In order to solve this issue, stepwise regression can be used instead of forward selection as it tends to be less greedy and fairer. Because forward selection by definition chooses predictors based on their entry p-values, it does not eliminate predictors based on their statistical significance in the broader model (no elimination based on exit threshold or in other words "in the context of the model").

4 Analyzing the Titanic Dataset

4.1 Logistic vs Linear Regression

Linear Regression: It is used to predict values of an outcome based on a linear relationship between the independent and dependent variables. For example, predicting the price of a stock, grade of a student, and weather temperatures.

Logistic Regression: It is used for problems in predicting the probability of a categorical outcomes, bounding the output between 0 and 1. For example, predicting the probability of survival in a car crash, the probability of a recession, and probability of finding a job. Usually, values in a linear regression model can be bounded by a logistic function to predict the probability of different outcomes rather than continuous values.

4.2 Probabilty of Survival

4.2.1 Steps

To find the probability of survival, the formula

$$\text{Probability of Survival} = \frac{\text{No. of survivors}}{\text{No. of passengers}}$$

Since survival in the dataset is 1 and death is 0, it can be calculated using the average by summing all 1s and dividing by the total number of datapoints.

4.2.2 Results

The probability of survival on the Titanic is approximately 38.2%

4.3 Probabilty of Survival by Age, Gender, and Passenger Class

4.3.1 Steps

To find the probabilities based on Age, Gender, and Passenger Class, categorizing the range of numerical values for age is useful. Instead of finding the probability of survival for each Age, they could be categorized as Infant, Child, Teen, Adult, Senior; this way, it becomes easier to visualize the data in a contained table. Infants are classified as younger than 5, Children are classified between 5 and 12, Teens are classified between 12 and 18, Adults are classified between 18 and 60, and Seniors are older than 60 (Table 8).

Infant	Child	Teen	Adult	Senior
<5	5-12	12-18	18-60	60>

Table 8. Age categories for Titanic passengers

4.3.2 Results

		Age Group				
Passenger Class	**Sex**	Infants	Children	Teenagers	Adults	Seniors
1	Female	0%	NaN	100%	97.37%	83.33%
	Male	100%	100%	50%	35.94%	6.67%
2	Female	100%	100%	87.5%	87.80%	NaN
	Male	100%	100%	0%	8.40%	16.67%
3	Female	65%	10%	60.71%	43.01%	100%
	Male	35.29%	33.33%	8.11%	16.18%	0%

Table 9. Probability of survival for the different age groups on the titanic. NaN signifies no people in the category.

A few observations, first, female survivals stand out relative to males in all categories quite significantly. From the scene in the Titanic, "For the time being, I shall require only women and children", thus it can be verified that they were scientifically accurate in the movie. Also, infants and children have high survival rates possibly due to the fact that they were accompanied by their mothers as they were being rescued. Across passenger classes, class 3 has the lowest survival rate among all age groups, except for senior women. Teenage, adult, and senior men on the other hand have the lowest survival rates also due to the women and children getting rescued first and the men being left behind.

4.4 Logistic Regression Model

4.4.1 Steps

The logistic regression model is one that is used to determine the probability of a certain outcome, where in this case it is either survival or death. The logistic sigmoid function, also called the squashing function, maps out the whole real axis [-inf, +inf] into a finite interval [0, 1] and it is given by

$$y(x) = \frac{1}{1 + e^{-x}}$$

where x is the dependent variable.

To train the logistic regression model, the goal is to use the sex, age, and pclass as the independent variables to predict the probability of survival on the titanic.

After building the model, it is appropriate and necessary to evaluate whether the variables are statistically significant and have a p-value < 0.05. If so, that means that

4.4.2 Results

Variable	**P-value**
sex	2.04×10^{-37}
age	9.83×10^{-7}
pclass	4.00×10^{-19}

Table 10. Sex, Age, and Passenger Class variables p-values in the final logistic regression model

All three variables sex, age, and pclass are statistically significant and are well below the threshold of 0.05. This means that the variables are relevant in predicting the survival rates on the titanic which are subsequently affected by sex, age, and passenger classes.

4.5 Model Performance

4.5.1 Steps

Model performance in a logistic regression can be measured by how many correct binary predictions it did. The correct predictions have to be made in both categories (0 and 1), and not only in predicting one of the measurements. If the model can only predict two outcomes that can either be right or wrong, the visualization of this scheme can be done using a confusion matrix.

True / Predicted	0	1
0	True Negatives (TN)	False Positives (FP)
1	False Negatives (FN)	True Positives (TP)

Table 11. Confusion Matrix Setup

If the data point was actually a death (or survival) and the predicted outcome was also a death (or survival), one point is counted in $\mathrm{CM}_{1,1}$ (or $\mathrm{CM}_{2,2}$). If the data point is not correctly predicted, it goes in either $\mathrm{CM}_{1,2}$ or $\mathrm{CM}_{2,1}$, depending on whether the actual data point is a death or survival, respectively in this case.

Then, to calculate the accuracy of the model, it is the ratio of how many correctly predicted values and the total number of predictions. It can be denoted using the following formula:

$$\text{Accuracy} = \frac{n(\mathrm{CM}_{1,1}) + n(\mathrm{CM}_{2,2})}{N}$$

where n is the mathematical notation for counting values, $\mathrm{CM}_{r,c}$ is confusion matrix with a row r and column c, and N is the number of data points measured.

4.5.2 Results

	Predicted (Deaths)	Predicted (Survivals)
Actual (Deaths)	107	17
Actual (Survivals)	18	68

Table 12. Confusion Matrix for

To calculate the accuracy of the model from the confusion matrix, add the values in diagonal [1,1] and [2,2] and divide by N (210 for the test set). Hence,

$$\text{Accuracy} = \frac{107 + 68}{210} \approx 83.33 \,\text{percent}$$

This means that the model was able to correctly predict 83.33% of the test data. However, for fairness, the test and training dataset was stratified to equally balance sample proportions so that the model is not training more or less categories and than it is being tested on. Moreover, the training size was 80% of the original data and then it got tested on 20% of the data.

Now, an accuracy of 83.33% could be satisfactory depending on the benchmark (previously discussed in this report). In general, 83.33% accuracy indicates that the model is capable of generalizing and not overfitting (memorizing the training dataset).