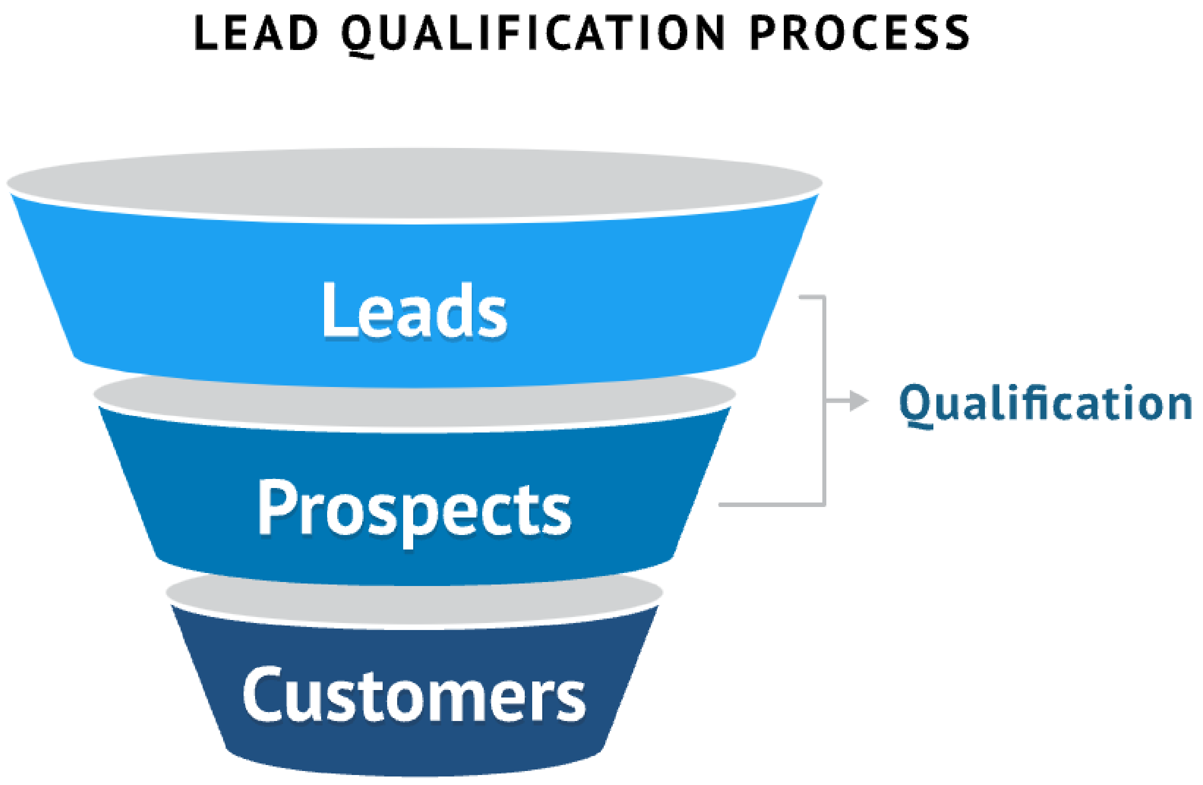
Logistic Regression with Stepwise Selection for Predictive Lead Scoring

Tawfik Fadzil

Lead Qualification



Lead qualification refers to the process of determining which potential customers are most likely to make an actual purchase.

An integral part of the sales funnel, which often takes in many leads but only converts a portion of them.

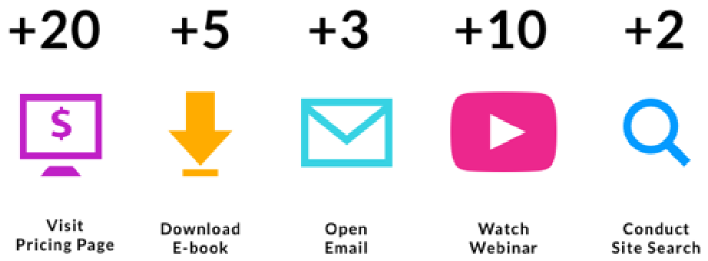
Lead Scoring

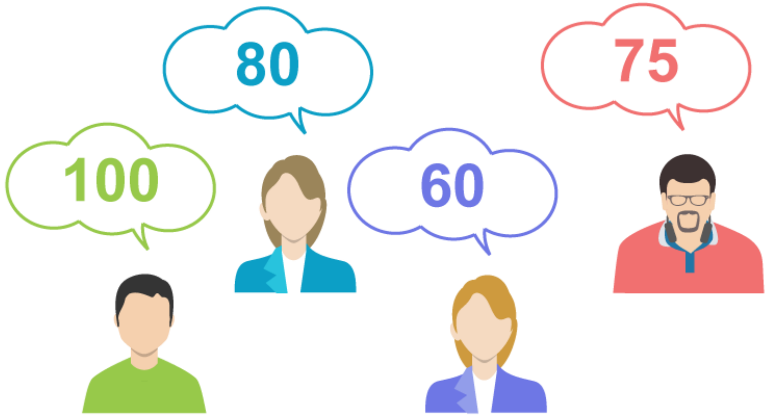
- Lead scoring is a popular methodology used by marketing and sales team to determine how likely their leads are to buy by means of a scoring system (often 1-100).

- Points are determined based on various attributes such as demographic characteristics and behavioral features.

- Once a lead reaches specific point total, they are considered a hot prospect and will be contacted by sales team.

- Leads with lower scores will be treated with proper marketing content and follow ups.





Manual vs Automated Lead Scoring

Nurture
Send to sales
Will never close
54
75
15
tye

<u>Manual Lead Scoring</u>

Assigning of score to the lead is manually done.

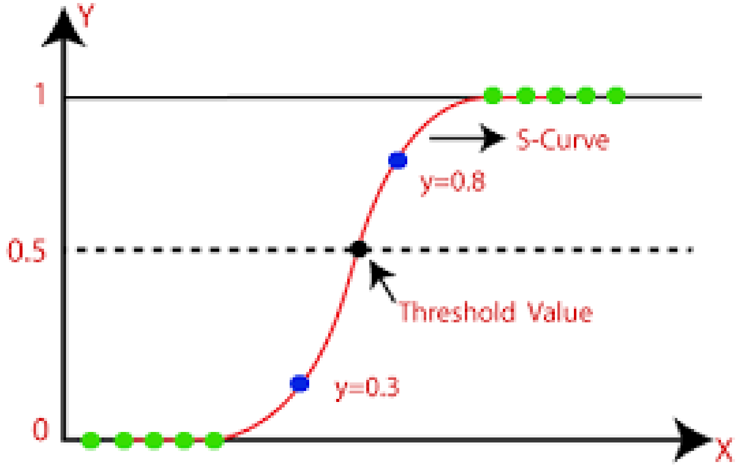
- Based on intuition and gut feeling.
- Not data driven.
- Prone to human error.

<u>Automated Lead Scoring</u>

Used machine learning to rely on the past data to understand the pattern and make prediction of likelihood to convert.

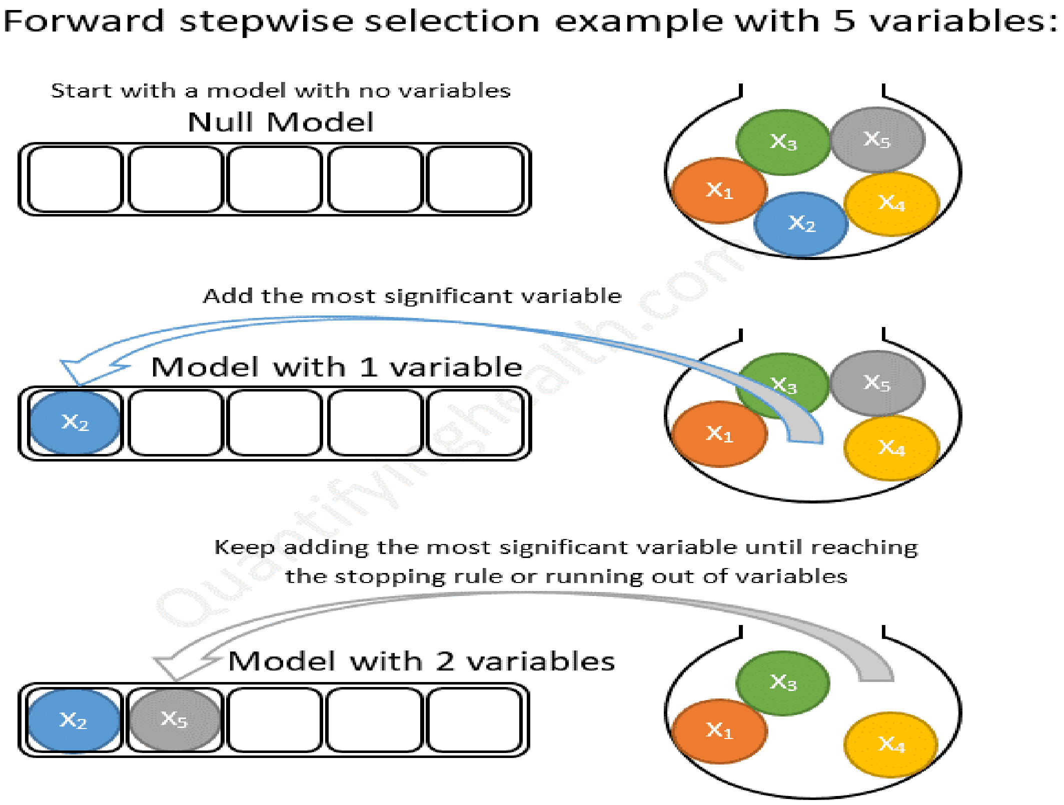
Logistic Regression

- Logistic regression (LR) is a common technique used in binary classification problems.
    - Widely used due to the simplicity of the model and easy interpret.
    - Can directly generate probability of a binary outcome.
    - Many researchers used LR as a benchmark for what a linear classification algorithm could achieve compared to other complex, non-linear machine learning algorithm.
- Well suited to study on the relationships between a categorical outcome variable and one or more categorical or numerical predictor variables.
- In many cases, certain predictors are not significant and not useful in predicting the dependent variable. For large number of predictor variables, selecting subset of important predictors manually can be very time consuming.
- A stepwise selection methods is useful to select and identify a useful subset of the important predictors, as well as the appropriate model.



Stepwise Selection Method

1) Forward Stepwise



Forward stepwise selection example with 5 variables:

Start with a model with no variables

Null Model

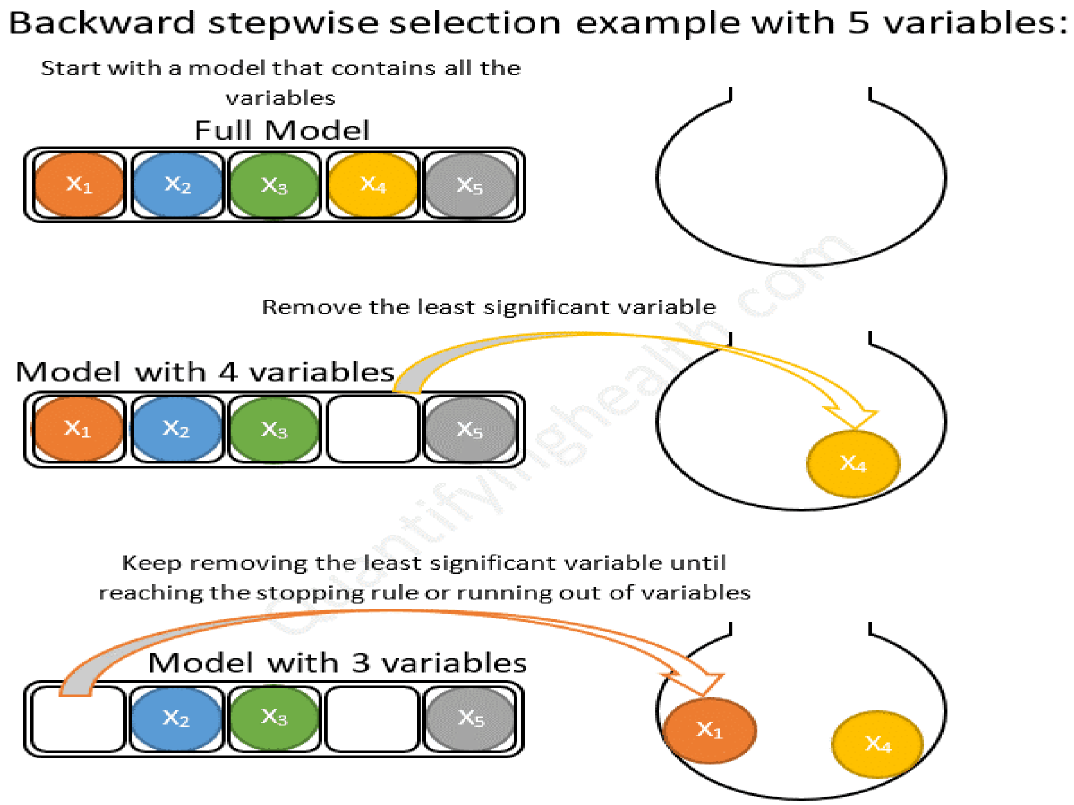
Add the most significant variable

Model with 1 variable

Keep adding the most significant variable until reaching the stopping rule or running out of variables

Model with 2 variables

2) Backward Stepwise



Backward stepwise selection example with 5 variables:

Start with a model that contains all the variables

Full Model

Remove the least significant variable

Model with 4 variables

Keep removing the least significant variable until reaching the stopping rule or running out of variables

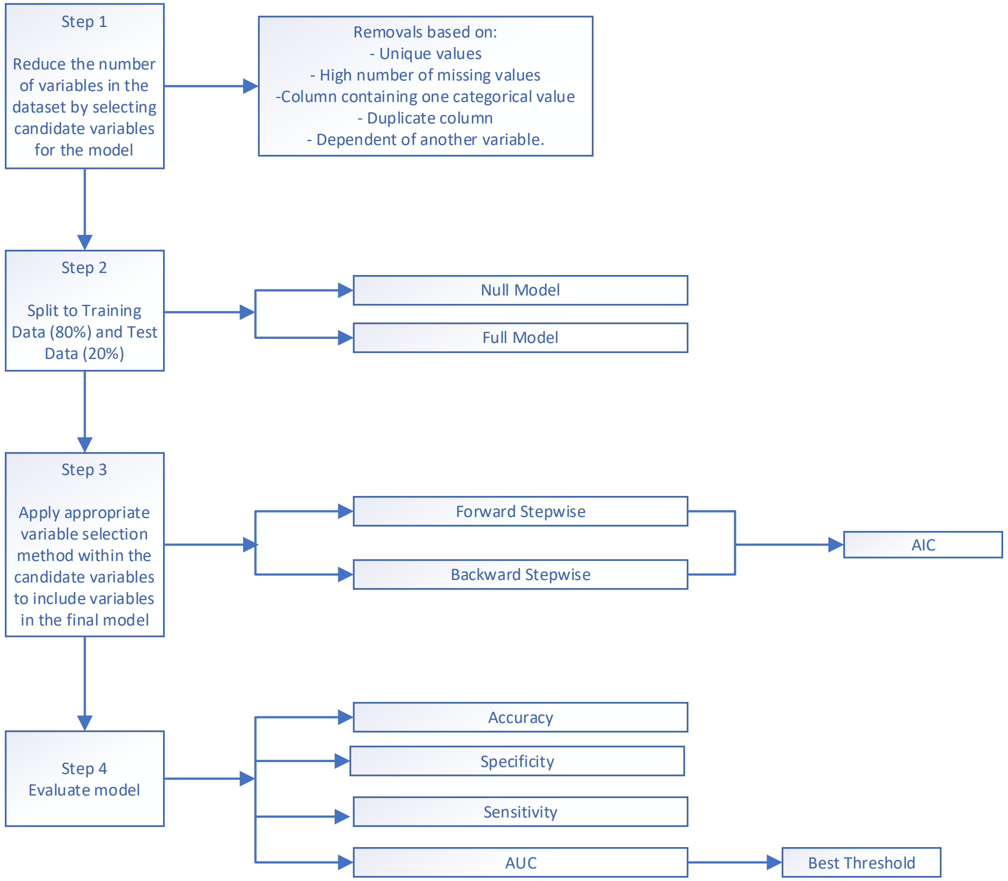
Model with 3 variables

Objectives

- To develop a predictive model for lead scoring using logistic regression.
- Identifying and selecting the major factors that influence the likelihood of successful sales conversion.
- Developing a parsimonious model which are simple but with great explanatory predictive power.

Methods

- Data Cleaning & Preprocessing
- Split Training Data (80%) and Test Data (20%)
- Run Logistic Regression for full model and null model.
- Using Forward and Backward Stepwise Regression based on AIC selection criteria.
- Evaluate models through Accuracy, Specificity, Sensitivity and AUC.
- Determine the best threshold.



Step 1

Reduce the number of variables in the dataset by selecting candidate variables for the model

Removals based on:
- Unique values
- High number of missing values
-Column containing one categorical value
- Duplicate column
- Dependent of another variable.

Step 2

Split to Training Data (80%) and Test Data (20%)

Null Model

Full Model

Step 3

Apply appropriate variable selection method within the candidate variables to include variables in the final model

Forward Stepwise

Backward Stepwise

AIC

Step 4
Evaluate model

Accuracy

Specificity

Sensitivity

AUC

Best Threshold

Data Cleaning & Pre-Processing

- Raw Dataset – 36 predictor variables with 9204 observations
- Dependent variable – Converted (Won or Lost)
- Reason of removals:
    - Columns containing only unique values (Lead number & Prospect ID)
    - Columns that have too many missing values.
    - Columns that only have one categorical value.
    - Duplicate columns.
    - Variable that is dependent of another variable (Examples include "Specialization column which provides details of "occupation" variable).
- Final Dataset – 11 predictor variables with 3502 observations.

Data Transformation

Three Types of Data:

- Nominal
- Ordinal
- Continuous

Nominal Variables

- Plot Pareto chart.
- Categorical values with low number of observations are grouped together as “others”.
- Transformed into unordered factor data type.

Ordinal Variables

- Change to numerical values to indicate the order of the categorical values.
- Transformed into an ordered factor data type.

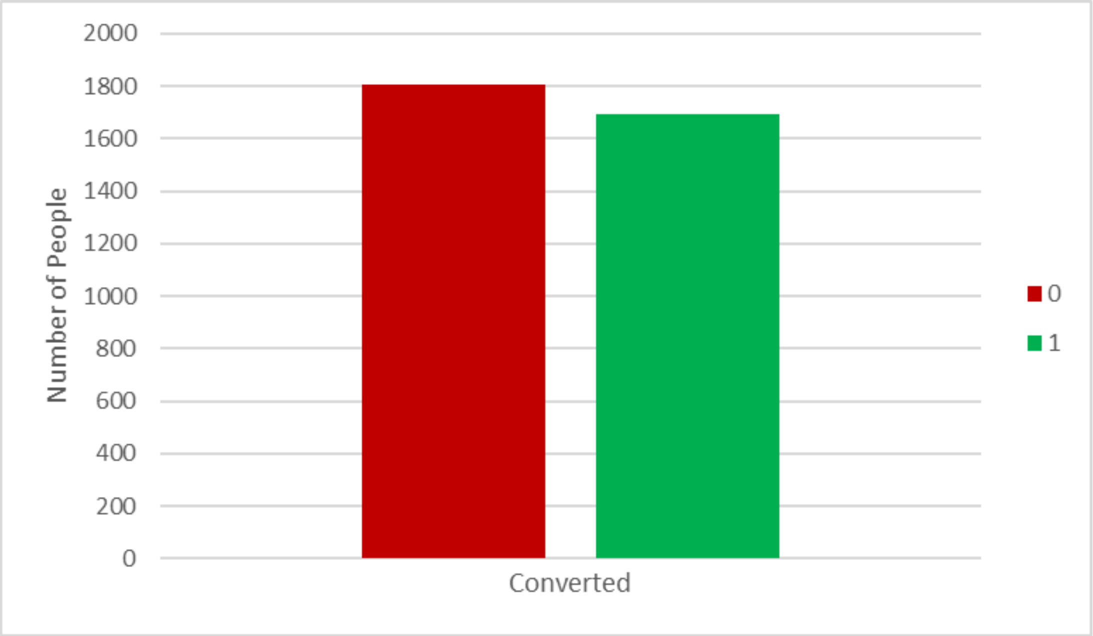
Continuous Variables

- Boxplot and descriptive statistics are used to understand the distribution of the data.
- Winsorising technique used to handle outliers

Converted (Won and Lost)

- “0” – Lost Deals (51.6%)
- “1” – Won Deals (48.4%)

- Distribution of the outcome is balanced.



Stepwise Selection

Forward Stepwise

	Step	AIC
1	Null Model	3880.611
2	+ Lead Quality	2840.992
3	+ Activity Index	2652.873
4	+ Total Time Website	2504.556
5	+ Lead Origin	2266.536
6	+ Lead Source	2209.278
7	+ Last Activity	2166.435
8	+ Current Occupation	2134.733
9	+ Total Visits	2131.443
10	+ Page Views Per Visit	2125.620

Backward Stepwise

	Step	AIC
1	Full Model	2130.907
2	- Profile Index	2127.095
3	- Free Copy	2125.620

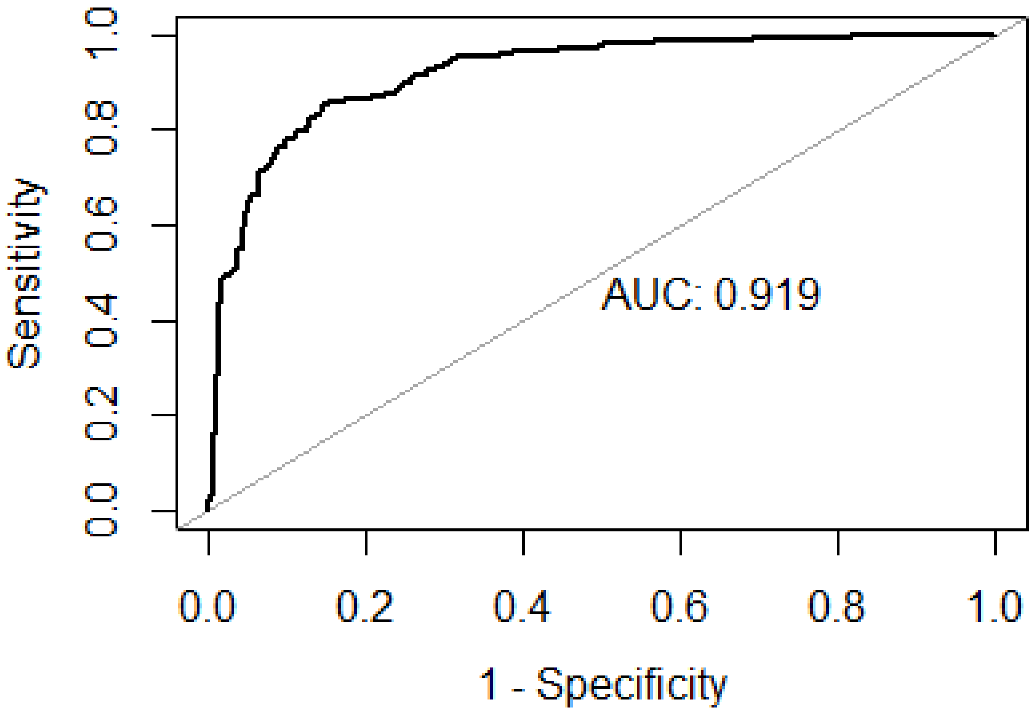
- Both forward and backward stepwise method produced the same results.
- Reduced from 11 variables to 9 variables.

Accuracy, Sensitivity and Specificity

	Accuracy	Sensitivity	Specificity
Full Model	85.02%	84.33%	85.71%
Stepwise Model	84.59%	84.05%	85.14%

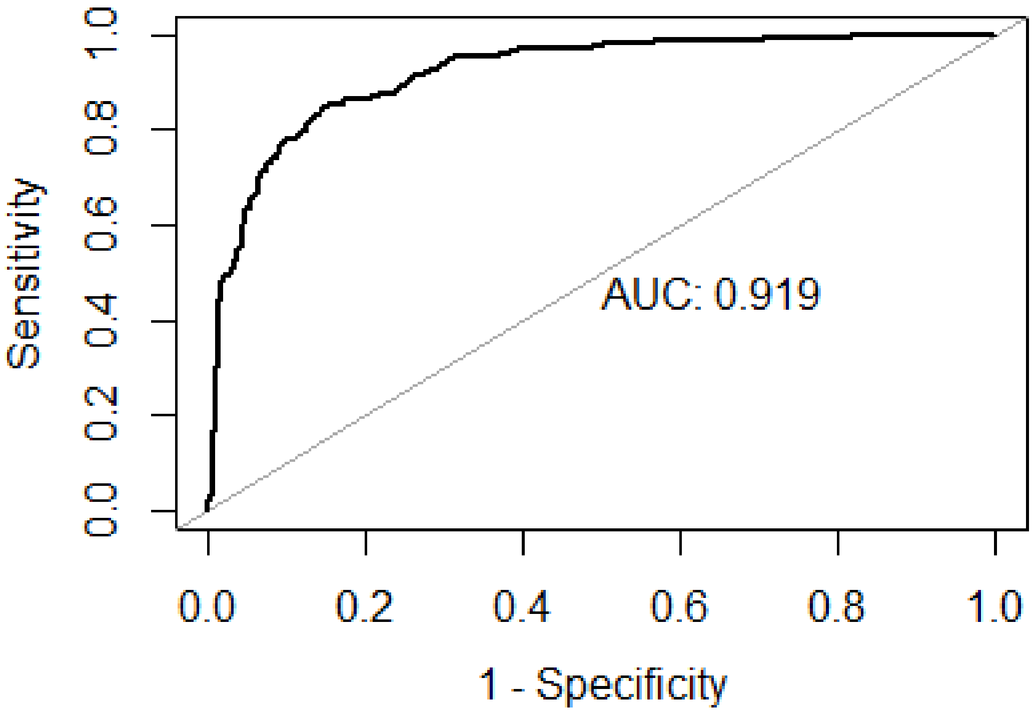
AUC-ROC Curve and Optimal Threshold

Full Model



Optimal threshold = 0.5034576

- Stepwise Model



Optimal threshold = 0.5230785

Conclusions

- Reached the same results when using forward or backward stepwise methods.
- Reduced from 11 to 9 predictor variables.
- Reached a simpler model without compromising much of the accuracy.
    - 85.02% - full model
    - 84.59% - stepwise model
- Very high AUC = 0.919 – Models have good performance in distinguishing between the positive and negative classes.
- Best threshold at around 50%
    - 50.35% - full model
    - 52.31% - stepwise model