Submitted to *INFORMS Transactions on Education*
manuscript (Please, provide the mansucript number!)

An Interactive Spreadsheet Model for Teaching Classification using Logistic Regression

Bahman Naderi
Amazon Web Services, Seattle, USA,

Vahid Roshanaei, Opher Baron
Operations Management & Statistics, Rotman School of Management, University of Toronto, Toronto, Canada,
Vahid.Roshanaei@rotman.utoronto.ca, Opher.Baron@rotman.utoronto.ca

We present an interactive spreadsheet that supports teaching essential concepts in classification and the use of Logistic Regression (LoR) model for binary classification. The spreadsheet allows to *interactively* compare linear regression with LoR along common performance measures. This comparison demonstrates the benefits of LoR by integrating computation with visualization. Students will reinforce concepts like probabilities, maximum likelihood estimation (MLE), and the usage of likelihoods to determine optimal parameters for the LoR. We then discuss classifications using the LoR while adjusting its decision boundary (DB). Students will learn how to convert assigned likelihoods into classification using the DB; impact classification outcome by varying DBs; designate predictions as True Positive, True Negative, False Positive, or False Negative; and determine the classification accuracy based on various performance measures, including Sensitivity, Specificity, Precision, Negative Predictive Value, F1 and F2 Scores, the Receiver Operating Characteristics curve, and lift/decile charts. These measures are depicted and dynamically adjusted when the DB changes. We also reiterate the usage of these measures in the context of cross validation and imbalanced data sets. We provide a case study that implements LoR with multiple predictors and an option for teaching the details behind MLE. We discuss pedagogical aspects of this spreadsheet.

Key words: Classification, Logistic Regression, Interactive Spreadsheet, Performance Measures, Cross-Validation, Balanced and Imbalanced Datasets, Maximum Likelihood Estimation, Excel Solver

1. Introduction

Logistic Regression (LoR) (see e.g., Chapter 4 of James et al. (2014)) is probably the most popular classification technique with prevailing applications (other common techniques are linear discriminant analysis and K-nearest neighbourhoods). Almost all textbooks related to statistical learning, data science, and business analytics teach LoR, typically, as the first classification

method. Therefore, LoR serves as the foundation for the comprehension of other classification techniques. As such, enhancing comprehension of LoR will culminate in a concrete assimilation of other classification techniques. We present an interactive and self-contained spreadsheet model that allows students to effectively grasp the statistical concepts related to the LoR classification model and understand common performance measures used for classification with balanced and imbalanced data.

LoR has been recently applied to (i) model the probability of a customer to patronize a firm among the existing firms as a function of sum of travel time and queuing delay (Dan and Marcotte, 2019), (ii) provide real-time estimation of portfolio risk measures and classification of portfolio risk levels (Jiang et al., 2020), (iii) estimate the enrollment probability of each applicant in the Korea Advanced Institute of Science and Technology (Kim et al., 2019), (iv) assist men in deciding whether or not to schedule a prostate cancer screening exam (Liberatore et al., 2009), and (v) teach students how to determine the relationship between point spreads and the probability of winning a game using data from the National Football League (Huggins et al., 2020). These applications of the LoR method demonstrate that this concept is not limited to a specific discipline and it ranges from medical to business to engineering, among others. Kopcso and P. (2018) present new ways to frame classroom discussion around the business value of models in data science, predictive analytics, and management science classes. The authors explore an example that integrates predictive analytics with prescriptive models in that predictive analytics are used to determine the inputs to the prescriptive models. The authors provide a review of predictive and prescriptive methods and how they map to business problems. In Table 1 of Kopcso and P. (2018), the authors establish correspondence between common business questions and analytics techniques. The authors propose LoR (among other techniques) to particularly address three business questions: (i) "Determine the impact of various factors on an output variable of interest", (ii) "Assign a score (e.g., likelihood, ranking) to an observation", and (iii) "Classify an observation (customer, loan, transaction, etc.) into a category". We discuss these questions in Sections 3.3, 3.4, and 3.8, respectively. Those interested in integrating Machine Learning (Predictive Analytics) into Operations Research (Prescriptive Analytics) curriculum are referred to the paper of Boutilier and Chan (2021). Our interactive spreadsheet model can support teaching the integration of predictive and prescriptive analytics.

In another recent study, Brusco (2021) develops two Excel workbooks that help with the understanding of (i) the maximum likelihood estimation problem; (ii) the process for testing the significance of LoR coefficients; (iii) different methods for model selection to avoid over-fitting; and (iv) the measurement of relative predictor importance using all possible subsets. We include topics (i) - (iii) in the Appendix and Sections 3.6 and 3.7, respectively. Our spreadsheet model

complements the works of Kopcso and P. (2018) and Brusco (2021) by incorporating more subjects that are taught using *interactive visualization* that are better suited for teaching to commerce and MBA students. Furthermore, we provide in-depth discussion of relevant performance measures for balanced e.g., Area Under the Curve (AUC), and imbalanced classification problems, e.g., F_β scores, Negative Predictive Value (NPV), Positive Predictive Value (PPV), decile/lift charts, which are common in practice. Other authors developed spreadsheet models for teaching analytics models to students. The interactive spreadsheet model of Erkut and Ingolfsson (2000) show students what error the linear regression minimizes. The authors used a *square shape* to illustrate the amount of Squared Error of each observation that emanates from the squared vertical distance of each observation and the fitted line. Huggins et al. (2020) developed a spreadsheet-based case study using real data from the National Football League to teach students how to establish the relationship between points spreads and the probability of winning a game by virtue of an LoR model. To tackle this case study, students require a combination of several key Microsoft Excel functions, including Pivot Tables, trend lines, and Solver.

Summary of contributions. The pedagogical and technical contributions of this paper are:

1. We develop a *comprehensive* spreadsheet model that allows instructors to *interactively* teach various aspects of classification to students in Science, Business, Analytics, Statistics, and Engineering using LoR. Our spreadsheet model allows students to grasp key concepts in LoR by using built-in interactive features in our spreadsheet model and to acquire deep insights into how various components of an LoR model are interconnected. Our spreadsheet includes the following sheets: (a) Read me, (b) Data, (c) Line-Curve, (d) Likelihood (Errors), (e) Decision Boundary, (f) Performance Measures, (g) Model Selection (h) Cross Validation, (h) a Case Study, and (i) Optional MLE Explained.[1] While some of these subjects may be separately taught in different lectures (e.g., cross validation and MLE), our spreadsheet model provides an overarching bridge among these subjects within a standalone lecture, similar to the lectures in the "Statistics for Managers" taught at the Rotman School of Management.[2] We present pedagogical highlights of each sheet in Section 3, where we explain them in greater details.

2. Our spreadsheet model improves the interaction between students and the instructor and between students and the subject matter. Once the instructor covered the learning objectives and pedagogical highlights for each concept in each sheet, students can manipulate different cells in the spreadsheet model and observe how performance measures are impacted by changes in the

[1] MLE stands for Maximum Likelihood Estimation.

[2] We cover sheets (a)-(h) in a two hours lecture and sheet (i) in another hour. We believe the teaching of all sheets is necessary for the development of a deep understanding of the LoR model and classification techniques for students who major in analytics. Instructors should use their discretion in choosing the teaching material depending on their learning objectives.

nonlinear curve that is fitted to data. We maintain that *interactivity* and *visualization* are very powerful *pedagogical tool* that can be used to (i) reinforce abstract concepts (like probabilities, likelihoods, and accuracy) and (ii) maximize students' engagement, culminating in a firmer and lasting grasp of the LoR model and classification.

3. The Case Study spreadsheet allows instructors to summarize and reiterate the learning objective associated previous sheets and teach students the LoR model with *multiple predictors*. While we have instantiated this sheet with a specific *imbalanced* dataset, instructors can use their own dataset to teach LoR with multiple predictors. Plugging data in this sheet will automatically calculate and generate relevant performance measures. In addition to being flexible in accepting new data, this sheet enables instructors to easily assign different cases to students. Moreover, the Case Study sheet provides instructors with excellent pedagogical opportunities to *interactively* demonstrate (i) how the determination (and change) of the "training proportion" of the dataset impacts the performance of the LoR model on both the training and test sets, (ii) how the "Excel Solver" is used to optimize the coefficients of the LoR model in the presence of multiple predictors and varying values of the training proportion, (iii) how changes of the DB impact the quality of classification, and (iv) which performance measures are best suited for imbalanced datasets. This sheet enables instructors to focus on the interplay among these parameters and demonstrate the back-and-forth nature associated with effective calibration of an LoR model's coefficients.

We organize this paper as follows. Since logistic regression and linear regression both fall under generalized linear models, we present an overview of the linear regression in Section 2 and then establish the relationship between the linear regression and the logistic regression. In Section 3, we further present pedagogical highlights of the concepts depicted on each sheet. We conclude in Section 4.

2. The relationship between linear and logistic regression

We present the fundamentals of linear and logistic regressions and explain why the latter is more adequate for classification. We focus on a *binary* classification problem (classification to two classes) based upon a *single* predictor, i.e., univariate regression.

There are many applications where one is interested in knowing the relationship between a predictor and a *qualitative response*, including gender, nationality, eye colors, employment status, etc. In such situations, the user wishes to *classify objects into well-defined groups*. In the context of such classification problems, it is natural to think of a probabilistic model that assigns a probability of belonging to a certain class to each data point. Focusing on classifications for two groups, we denote the two classes as class 0 and class 1 (i.e., the response variable Y can be either 0 or 1), we look for a model that predicts the probability of $Y = 1$ for a given X, i.e., $p(X) = Pr(Y = 1|X)$, where, obviously, $p(X) = 1 - Pr(Y = 0|X)$.

2.1. Linear regression for prediction

We first discuss the shortcomings of applying a linear regression to solve a classification problem with a binary response 0 and 1. Univariate linear regression uses the linear function

$$Y = \beta_0 + \beta_1 X \tag{1}$$

to find the best line that goes through a set of data points using the (ordinary) *least squares* method. Finding the best line entails optimizing the values of the *intercept*, β_0, and the *slope*, β_1, that minimizes the sum of squared errors. The optimized linear regression line is written as

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X. \tag{2}$$

The slope of the above model can be interpreted as the average amount of increase (or decrease if the sign of $\hat{\beta}_1$ is negative) in the response variable Y as we increase the predictor variable, X, by one unit. The intercept can be interpreted as the value of Y when the predictor X is at zero value. Note that this model assumes that the response variable Y is *quantitative* and the value of Y is obtained by setting the value of $X = x$ where x is a specific value of the predictor.

An important limitation of (1) is that it is not a natural model for classification as it does not capture the probability assigned to each predictor value X directly. To address this limitation consider the following linear regression model for classification:

$$p(X) = \beta_0 + \beta_1 X. \tag{3}$$

However, this linear model may produce $p(X) > 1$ or $p(X) < 0$ that cannot be intuitively interpreted from a probabilistic perspective.

2.2. Logistic Regression for classification

An alternative method for classification is logistic regression (LoR). In LoR, we use the following logistic function:

$$p(X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}} \tag{4}$$

that unlike linear regression, finds the best *nonlinear S-shaped fit* to the data using the Maximum Likelihood Estimation (MLE) method.[3] Note that logistic function produces an output of (0,1): when $x = -\infty$ the prediction in (4) would tend to 0 and when $x = \infty$ this prediction would tend to 1.

[3] Details of MLE are provided in optional sheet MLE Explained. Instructors may consider teaching this optional sheets if they want to provide more details as to how the coefficients of the LoR model are optimized using the MLE method.

In order to establish the relationship between the linear regression and the logistic regression, we define odds as $\frac{p(X)}{1-p(X)}$ and note that the odds can take any value in $(0, \infty)$. Based on the definition of odds, we rewrite Equation (4) as

$$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}. \tag{5}$$

We next take the logarithm of Equation (5) as follows:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X. \tag{6}$$

The left-hand side of Equation (6) is called the *log-odds* or *logit*, implying that logistic regression model, as captured in Equation (4), has a logit that is linear in X.

3. Interactive Spreadsheet Model for Classification

In previous section, we provided basic concepts in Linear and Logistic Regressions. In this section, we explain how our spreadsheet model allows students to interactively learn key concepts in classification and the usage of LoR model for classification.

Similar to the concept of error minimization in linear regression, the LoR model finds a *nonlinear Sigmoid function* that closely fits the data using the concept of Maximum Likelihood Estimation (MLE). Our interactive spreadsheet helps visually reinforce the concept of MLE as it allows students to see how different parameters of the nonlinear function impact the amount of Maximum Likelihood.

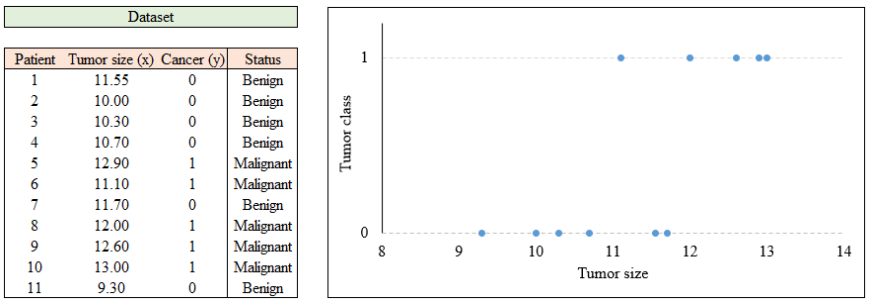
The Read Me sheet is the first in our spreadsheet. In this sheet we explain how instructors should teach each sheet of the spreadsheet model. This sheet also highlights crucial pedagogical points for the other sheets.

3.1. Data sheet

In this data sheet, we provide data that enables instructors to investigate various aspects of LoR. The dataset table consists of data on 11 patients, their tumor size that would serve as the predictor, and the final **binary** diagnosis for these tumors that is the response we are trying to predict. Note that $y = 0$ if the tumor is benign (the tumor does not contain cancerous cells) and $y = 1$ if the tumor is malignant (the tumor contains cancerous cells). Our dataset and its plot are shown in Figure 1.

The binary classification in the data is based upon medical screening of tumors and other related diagnostics. These tests classify the tumor as either **benign** or **malignant**, denoted by 0 and 1, respectively. The goal of the classification is to learn the relationship between the tumor size and its status.

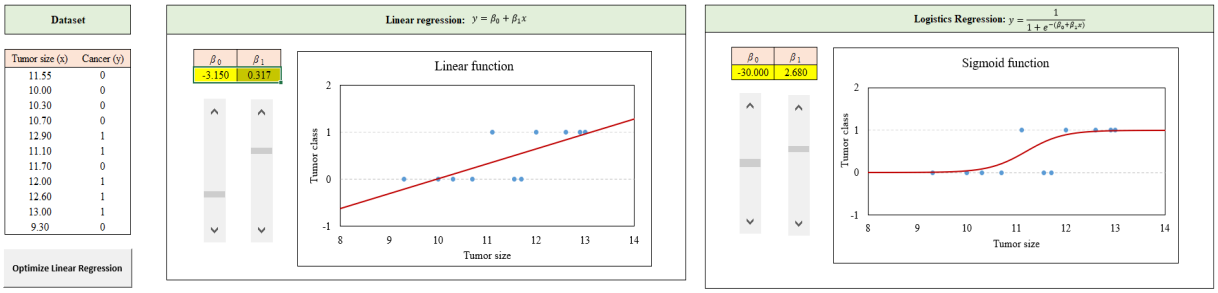
One pedagogical moment is to remind students that zeros and ones under column "Cancer (y)" are indeed "Benign" or "Malignant" tumors that by virtue of e.g., Excel function =IF(desired



Dataset

Patient	Tumor size (x)	Cancer (y)	Status
1	11.55	0	Benign
2	10.00	0	Benign
3	10.30	0	Benign
4	10.70	0	Benign
5	12.90	1	Malignant
6	11.10	1	Malignant
7	11.70	0	Benign
8	12.00	1	Malignant
9	12.60	1	Malignant
10	13.00	1	Malignant
11	9.30	0	Benign

Figure 1 Visualization of the cancer dataset



(a) Linear fit to the data (b) Nonlinear (Sigmoid) fit to the data

Figure 2 Linear and nonlinear fits to the data

cell= "Malignant",1,0) have been converted into zeros and ones, respectively. The other point is to remind students of the obvious fact that y-Axis ranges between 0 and 1 for any observation of X, which is the tumor size. This basically indicates that any tumor, regardless of its size, is either class 0 or 1.

3.2. Line-Curve Sheet

This sheet includes two interactive figures that support classifications of our data using linear (Figure 2(a)) and logistic (Figure 2(b)) regressions. The bars on the left of each figure can be used to change the functions and observe their fit to data. The "**Optimize Linear Regression**" button is used to show optimal coefficients of the linear regression (i.e., the coefficients that minimize the sum of square errors of the linear regression). Observe that in contrast to the linear regression that leads to y values that are not necessarily between 0 and 1 for some tumor sizes, the LoR leads to y values between 0 and 1 for any tumor size.

Assuming that students are familiar with linear regression, we first allow students to fit a *straight line* to data points by playing around with parameter values of the linear regression model (1): intercept (β_0) and slope (β_1), and interactively see how the line is fitted to the data. Students can press button "Optimize Linear Regression" to see the pre-optimized values for parameters

intercept β_0 and slope β_1[4]. Having found these coefficients, student can observe that the linear regression leads to probabilities outside of [0,1] for some X inputs. This is the main reason why linear regression is not frequently used for classification. The instructor can also let students know that Excel functions "**Intercept**" and "**Slope**"[5] can be used to determine the values of β_0 and β_1, respectively.

One pedagogical highlight is that the instructor accentuates on the optimized coefficients of the linear regression in the context of probability assignment. For instance, the instructor can draw students' attention to the optimized β values ($\hat{\beta}_0 = -3.150$ and $\hat{\beta}_1 = 0.317$ state that for $X = 0$ and $X > 14$, the optimized linear function produces $P(Y = 1|X = 0) = -3.150$ and $P(Y = 1|X = 14) > 1$, respectively, none of which has a natural probabilistic interpretation. In fact, in Figure 2(a), we can easily see that for some predictor X values the response variable, $P(Y = 1|X)$, is outside the [0,1] range that is required for a probabilistic interpretation. In contrast, under any β values the logistic regression leads to prediction values that are in (0,1) for any size of tumor (see Figure 2(b)).

After highlighting to students the shortcoming of the linear regression, the instructor could manipulate the β coefficients of the logistic regression and ask students to do the same. This experimentation will help students develop a good understanding of what combination of LoR function's parameters leads to a nonlinear curve fit that is closer to data points (see Figure 2(b)). At this point, the instructor can highlight four pedagogical points regarding the LoR model:

1. For any values of β_0, β_1, and X, the logistic function (4) returns $p(X) \in (0, 1)$. The logistic function (4) gets extremely close to boundaries $y = 0$ and $y = 1$, but it never produces $p(X)$ exactly equal to 0 and 1;

2. For any β values the associated nonlinear curve will assign a probability of being malignant to each tumor with size $X = x$ and these probabilities can be seen by hovering the mouse on top of the red curve line near each X in Figure 2(b);

3. Remind that for the Linear regression model, regardless of the value of X, if β_1 is positive then increasing X will increase $p(X)$, and if β_1 is negative then increasing X will decrease $p(X)$; and

4. Explain that for the LoR model, increasing X by one unit will change log(odds) by β_1 or equivalently multiply the odds by e^{β_1}. For example when $\beta_1 = 2.43$, the log(odds) increases by 2.43 if we increase X by one unit, i.e., the odds are multiplied by $e^{2.43}$.

[4] These optimized parameters have already been obtained via the OLS technique.

[5] These two functions accept two vectors of data Y and X to determine the value of the intercept and slope.

3.3. Likelihood sheet

In this sheet, depicted in Figure 3, students will learn (i) how the logistic function produces **probabilities** and **likelihoods** and how the likelihood is used to optimize the β_0 and β_1 parameters (ii) how to differentiate between the concept of probability and likelihood, and (iii) the concept of **Maximum Likelihood** that results from the optimized nonlinear curve fitted to data points.[6] This sheet allows students to learn about the objective of the LoR function and differentiate between *Ordinary Least Squares* that minimizes error in the linear regression and the *Maximum Likelihood Estimation (MLE)* that maximizes the likelihood in the LoR model. This sheet allows students to interactively observe how choosing different parameter values impact the MLE value, $L(\beta_0, \beta_1)$ as depicted by the blue bar in the stacked bar chart, and to learn that $L(\beta_0, \beta_1)$ ranges between 0 and 1 with lower values of $L(\beta_0, \beta_1)$ indicating a poorer fit.

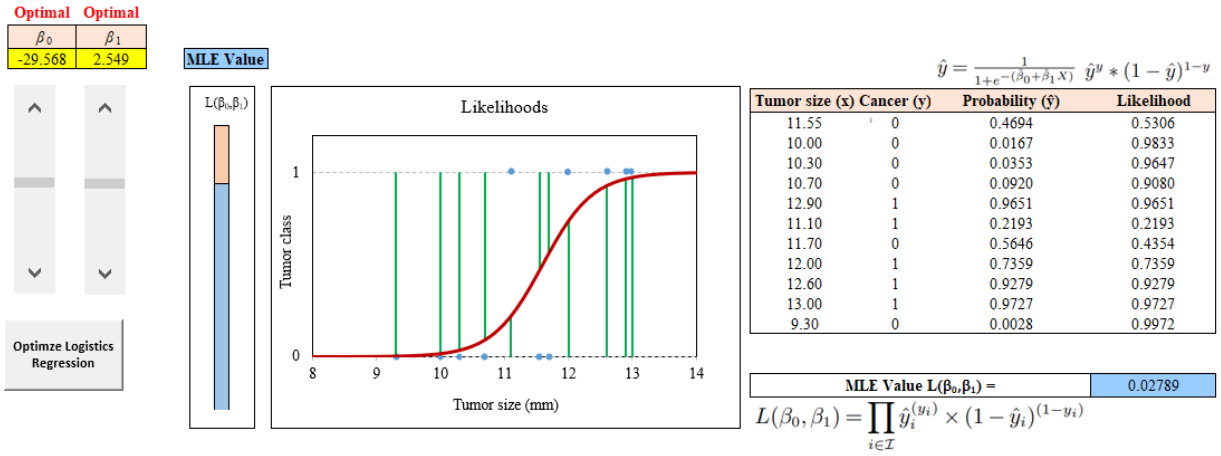


Figure 3 Probability, Likelihood, and Maximum Likelihood

Given a data point X (e.g., tumor size) and optimized parameters β_0 and β_1, the LoR model uses (4) to assign probability $\hat{y} = \frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1 X)}}$ that observation X belongs to class 1 (e.g., malignant) patients. (Note that we use $\hat{y}$ to denote the assigned probability $p(X)$ to observation X given the optimized values of β_0 and β_1, denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively.) The likelihood of these observations to be correct is the distance between this $\hat{y}$ and $1-y$. For example, the likelihood of a $y=1$ observation is simply $\hat{y}$. In our data set, the largest tumor with $X=13$ is malignant, i.e., $y=1$, so under the optimal $\hat{\beta}_0$ and $\hat{\beta}_1$ parameters, the LoR classifies this tumor as malignant with probability $\hat{y} = \frac{1}{1+e^{-(\hat{\beta}_0+\hat{\beta}_1 X)}} = 0.9727$ and the likelihood of this classification to be correct is 0.9727 as is captured by the green line of length 0.9727 that corresponds to the right most point ($X=13$) in Figure (3). Similarly, the likelihood of a $y=0$ observation is $1-\hat{y}$. In our data set, the

[6] The procedure for estimating Maximum Likelihood is presented in the "optional MLE Explained" sheet.

smallest tumor with $X = 9.30$ is benign, i.e., $y = 0$, so under the optimal $\hat{\beta}_0$ and $\hat{\beta}_1$ parameters, the LoR classifies this tumor as malignant with probability $\hat{y} = 0.0028$ and the likelihood of this classification to be correct is $1 - 0.0028 = 0.9972$ as is captured by the green line of length 0.9928 that corresponds to the left most point ($X = 9.3$) in Figure (3).

We can rewrite the likelihood of each observation X with associated probability $\hat{y}$ as

$$L_X(\beta_0, \beta_1) = \boxed{\hat{y}^y} * \boxed{(1-\hat{y})^{(1-y)}}. \tag{7}$$

Note that, for a $y = 1$ observation we can rewrite the likelihood using the formula included only in the first box: $\hat{y}^y$ because the formula in the second box is 1 (the power will be zero). Similarly, Equation (7) captures the likelihood of a $y = 0$ observation in that the formula in the first box is 1 (the power is 0) and the formula in the second box is the likelihood. So, the likelihood of any observation, as given by the green lines, is mathematically captured by (7). Moreover, the likelihood of *all* observations is captured by multiplying their independent likelihoods as:

$$L(\beta_0, \beta_1) = \prod_{i \in \mathcal{I}} \hat{y}_i^{(y_i)} \times (1-\hat{y}_i)^{(1-y_i)}. \tag{8}$$

There are multiple pedagogical points associated with this sheet:

1. Students learned that for any β_0 and β_1 the red S-curve in Figure (3) produces probability $\hat{y}$ for each tumor with size X to belong to class malignant, as in (4), $\hat{y} = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}}$. In case of a perfect classification, the likelihood of each data point is 1. For any other practical classifications, we wish to maximize this likelihood—getting it as close to 1 as possible for each data point.

2. Unlike in the sheet Line-Curve where students could not see whether the nonlinear curve was getting closer to data points as a function of changing β_0 and β_1, this sheet provides an adjustable stacked bar that allows students to interactively see whether the new choice of LoR coefficients improves the quality of the fit, which is measured by $L(\beta_0, \beta_1)$—higher values (blue color) indicate a better fit.

3. Having sufficiently played with these parameters, students can press the "**Optimize Logistic Regression**" button in which case the adjective "**optimal**", highlighted in red, will appear on top of the LoR coefficients' cells. Note that before pressing the "Optimize Logistic Regression" button, formula $L(\beta_0,\beta_1)$ provides sub-optimal likelihood estimations for observations. Once the button is pressed formula $L(\beta_0,\beta_1)$ yields the *Maximum* Likelihood Estimation (MLE) value.

4. The sheet explains the probability and likelihood measures and provides their mathematical formulas. For example, the formula for the calculation of likelihood of each point is provided in (7); therefore, given the optimal LoR model with $\hat{\beta}_0 = -29.568$ and $\hat{\beta}_1 = 2.549$, a data point $X = 11.55$, and $y = 0$, we obtain probability of 0.4694 ($\hat{y} = 0.4434$) and a likelihood of 0.5309.

5. Since LoR models calculates the probability of an observation belonging to class $y = 1$, higher values of $\hat{y}$ for an observation with $y = 1$ in our data set indicates a higher likelihood, i.e., a better fit of the LoR model. Similar logic holds for lower $\hat{y}$ for observations $y = 0$.

6. The objective function of the MLE, $L(\beta_0, \beta_1)$ is to maximize the product of the likelihood values for all observations. Moreover, as $L(\beta_0, \beta_1)$ is a product function of factors smaller than 1, it is decreasing with the number of observations.

Remark. *If the LoR model is presented in a 3-hour lecture, the instructor can teach sheet "MLE Explained" presented in the Appendix. This sheet develops an in-depth understanding of the method behind optimizing the coefficients of an LoR model (i.e., MLE) and is a suitable topic for technical students in Engineering, Science, and Analytics.*

3.4. Decision boundary sheet

Having found the optimized parameter values for the LoR model, one can use its output (probability) for classification. As we earlier stated, LoR assigns a probability of belonging to class 1, i.e., the tumor being malignant, to each observation.

In this sheet, we depict the as a horizontal black line superimposed on the optimized fitted LoR line. Students can change DB and dynamically observe how the classification of observations change due to the new value (see Figure 4). The DB helps convert probabilities into zeros and ones. Given a DB if $\hat{y} \leq DB$, the observation is assigned to class 0 and otherwise 1. For instance, if $DB = 0.5$ (which is the default DB for classification) and the probability of a data point is 0.469 (the probability of the first data point), we classify that data point as a benign tumor, i.e., we assign value 0 to this data point.

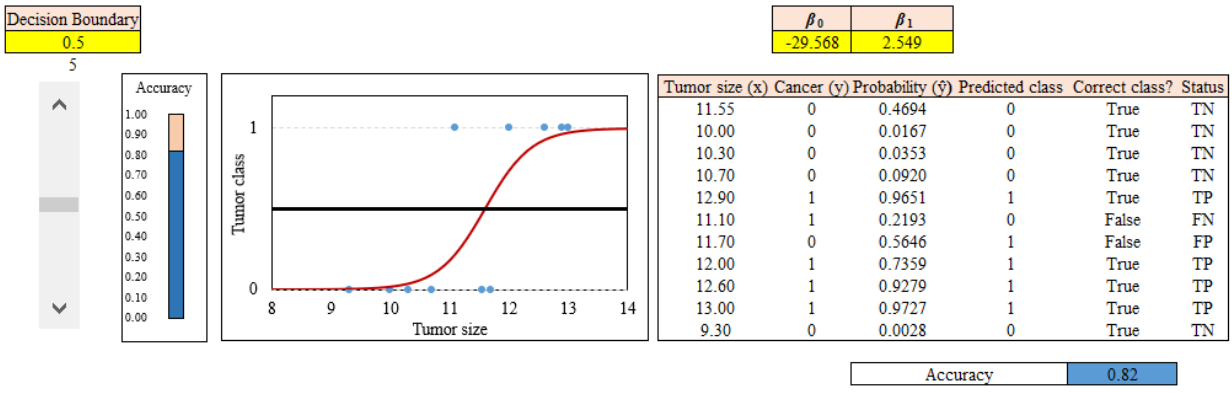


Figure 4 The impact of the value of DB on classification

Under the column "Correct class?" students can see whether the predicted class is correct for each data point (True) or not (False). (Excel returns "True" when the y value of each data point

under column "Cancer (y)" is equal to the value under column "predicted class" and "False" otherwise). Based on this comparison, we determine the **accuracy** of the classification. Accuracy of the classification is defined as the fraction of the total number of correctly classified data points divided by the number of all data points, which in our case is:

$$\textbf{Accuracy} = \frac{\text{Number of correctly classified patients (True)}}{\text{All patients (True + False)}}.$$

The value of accuracy is calculated in column M and is also illustrated in a stacked bar chart that is dynamically adjusted based on the DB.

In this sheet, we further introduce students to the status of classification: *True Positive*—the correct classification of a positive observation, *True Negative*—the correct classification of a negative observation, *False Positive* (aka **Type-1 error**)—the false classification of a negative observation as a positive one, and *False Negative* (aka **Type-2 error**)—the false classification of a positive observation as a negative one. We let TP, TN, FP, and FN denote the total number of data points classified, respectively, as *True Positive*, *True Negative*, *False Positive*, and *False Negative*. Then, the accuracy is defined a

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{9}$$

The DB is set by the decision maker based on the application and dataset to which the LoR model is applied.[7] In this sheet, we set the default DB to 0.5 and classified tumor types based on this value. Students can now play with the value of the DB and see which of patients' classification will change and understand its underlying reason. Any changes in classification of patients due to the change in the default DB is highlighted in red and students can focus on those cells to ascertain why the predicted class for each of these cells was changed and why this new classification is better or worse than the default setting.

The instructor can now set DB = 0.7. The highlighted cell in Figure 5 shows the data point that its class has changed due to new values of the DB.

A summary of pedagogical highlights for this spreadsheet is:

1. Explain that the "**Decision Boundary** (DB)" dictates assignment of each tumor into 0 and 1, i.e., benign and malignant tumors, respectively. Different DB yield different accuracy value for the LoR model.

2. Emphasize that the classification is performed based on *probabilities* and not likelihoods. The assignment of each observation is done by comparing the probability of an observation with the DB probability threshold. Observations with a (strictly) higher probability than DB are assigned as 1 and other observations are assigned as 0.

[7] More on this subject is covered in the case study sheet.

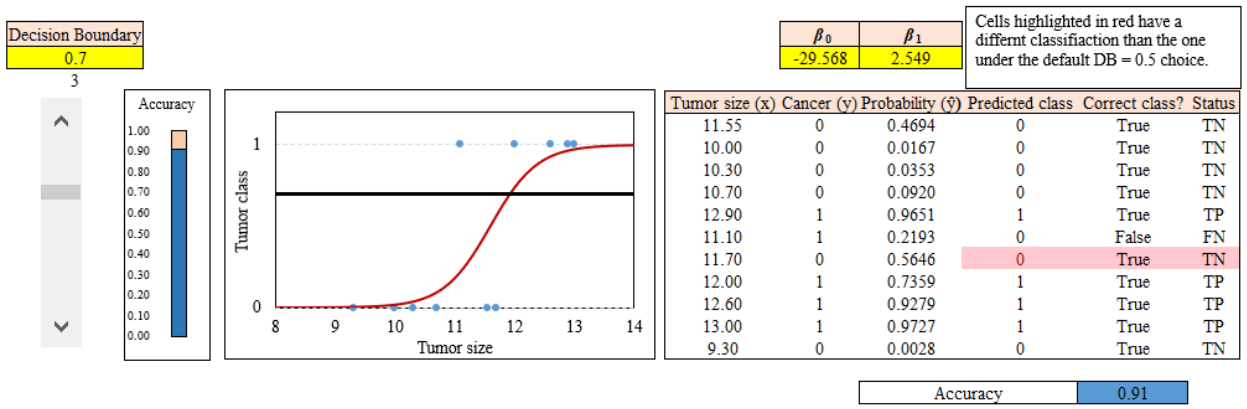


Figure 5 **The impact of the value of decision boundary on classification**

3. Note that the DB *does not* impact probabilities in any ways and probabilities are only determined by the LoR model coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$. Thus, the DB is a value that is determined *postmortem* by the decision maker and hence it is not optimized by the MLE method.

4. Higher values of DB culminates in fewer observations being classified as positives (smaller number of tumors being classified as malignant) and more observations being classified as negatives (more tumors are classified as benign). Thus, the choice of the DB depends on the trade-off between different performance of the classification.

3.5. Performance Measures Sheet

This sheet transcends the LoR model and applies to all classification techniques. Having students acquainted with the concepts of TP, TN, FP, FN, and accuracy the instructor can interactively teach other commonly-used performance measures for classification.

In Table 1, we present the **Confusion Matrix** that encompasses the aforesaid four rates.

Table 1 **Confusion Matrix: FP and FN (bold) rates indicates Type-1 and Type-2 errors, respectively.**

	Actuals	
Predictions	TP	**FP**
	FN	TN

We use this sheet (Figure 6) to introduce additional common measures for classification. The most important point that the instructor must raise is that all performance measures covered in this sheet are also influenced by the DB. This sheet allows students to explore the impact of various $\text{DB}_{\text{values}}$ on different performance measures of an LoR model. We divide these performance measures, based on the type of data, into two categories: (i) *balanced* dataset as presented in §3.5.1 and (ii) *imbalanced* dataset as presented in §3.5.2. We divide our performance measures for *balanced* datasets into two distinct groups: (i) recall (otherwise known as TP rate or as sensitivity) and

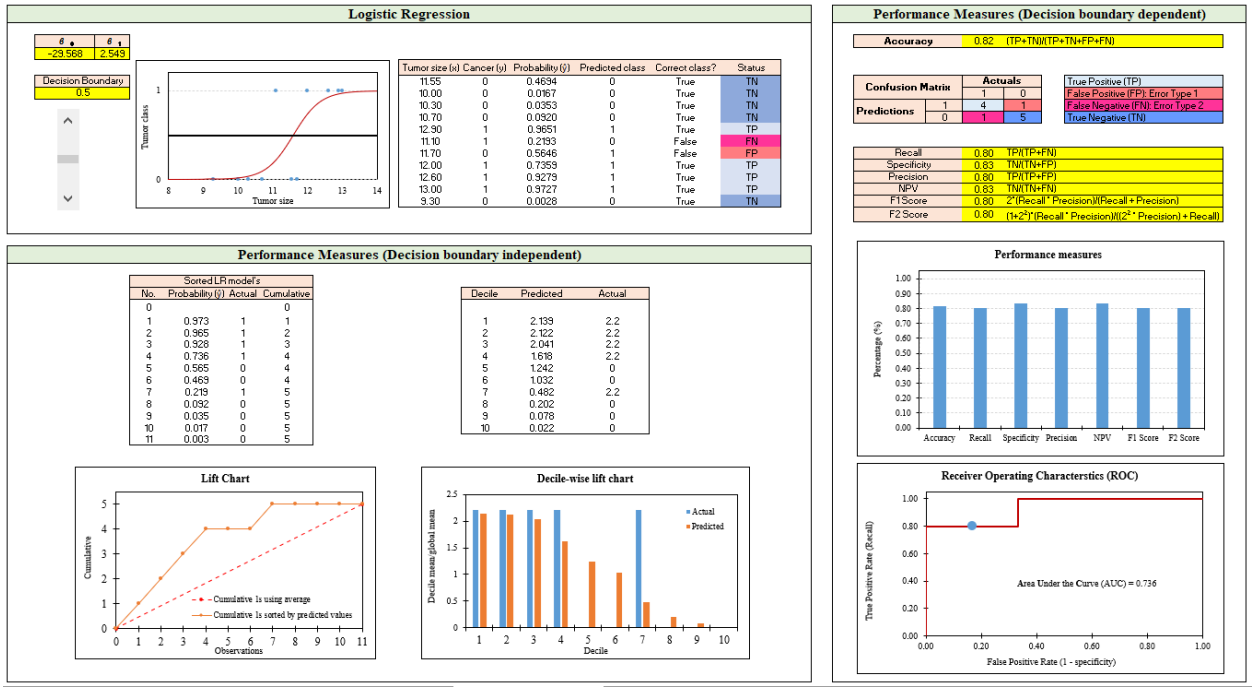


Figure 6 Classification performance measures

specificity (otherwise known as TN rate) that reveal *the distribution of classifications conditioned on the true outcome* and (ii) Positive Predictive Value (PPV, otherwise known as precision), and Negative Predictive Value (NPV) that *provide the probabilities of outcomes conditioned on the model's classification*. For *imbalanced* datasets, we introduce two performance measures F_β Score and lift/decile charts.

3.5.1. Performance measures for balanced datasets

(i) We first discuss recall and specificity. The formula for recall is:

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} = \frac{\text{Number of correctly classified positive observations}}{\text{All positive observations}}. \tag{10}$$

Recall is concerned with the correct classification of positive observations. In our example, Recall indicates what fraction of malignant tumors that has been correctly classified as malignant.

The formula for specificity is:

$$\text{Specificity} = \frac{\text{TN}}{\text{FP+TN}} = \frac{\text{Number of correctly classified negative observations}}{\text{All negative observations}}. \tag{11}$$

Specificity is concerned with the correct classification of negative observations. In our example, specificity indicates what fraction of benign tumors that has been correctly classified as benign.

The instructor can choose the value of the DB in such a way that it helps with minimizing one error type versus the other (or improving one performance measure versus another). To select a

DB to convert a probability model into a classification model, we must consider quantities like recall and specificity. Essentially, we aim to find a DB that maximizes both TP and TN rates in our **Confusion Matrix** as depicted in Table 1.

The trade-off between recall and specificity is depicted in our interactive **Receiver Operating Characteristics (ROC)** curve whose x- and y-axis are, respectively, the TP rate, i.e., recall, and the FP rate, i.e., 1- specificity. An ROC curve demonstrates the diagnostic ability of a binary classifier. For instance, in our data we have 5 patients with malignant tumors; the LoR model, at DB = 0.7, has assigned only 4 of these patients to the malignant category; therefore, the TP rate at $\text{DB}_{\text{values}} = 0.7$ is 0.80 (or 80%). At $\text{DB}_{\text{value}} = 0.7$, the FP rate on the x-axis of the ROC curve is zero (no benign patient classified as malignant) and the TP rate on the y-axis is 0.80. *For any DB that leads to a change in classification of observations, the trade-off in the ROC curve, as captured by the blue circle, is dynamically adjusted.* This blue circle is the only part in the ROC Figure that is impacted by the DB and this circle helps students visualize the impact of the DB on the FP and TP rates. While ideally one wishes to see the blue circle at the top left corner of the ROC curve, which indicates 100% TP rate and 0% FP rate. In application, maximizing both values simultaneously is not possible and one has to find the right *trade-off*. The instructor can choose $\text{DB}_{\text{value}} = 0.8$ (instead of 0.7) and show that the number of observations previously classified as TP decreased from 4 to 3, i.e., the new $\text{DB}_{\text{value}} = 0.8$ caused one of the previously correctly classified positive observations to be falsely classified as negative. Hence, for our sample, $\text{DB}_{\text{value}} = 0.8$ is worse than $\text{DB}_{\text{value}} = 0.7$ because it has deteriorated the TP rate without improving the TN rate. (Note that in our sample the performance of the classification is identical for DB=0.6 and DB=0.7.

(ii) Students have so far seen that the LoR model is not perfect and given this shortcoming, one might be interested in asking: given that the LoR model indicates a particular tumor is malignant, what is the probability that the tumor is actually malignant? This is the question that clinicians might ask after they see the result of the classification. This question is best responded by the PPV Measure:

$$\text{PPV} = \frac{\text{TP}}{\text{TP+FP}} = \frac{\text{Number of correctly classified positive observations}}{\text{All observations } \textit{classified} \text{ as positive}}. \tag{12}$$

A similar performance measure is the NPV:

$$\text{NPV} = \frac{\text{TN}}{\text{FN+TN}} = \frac{\text{Number of correctly classified negative observations}}{\text{All observations } \textit{classified} \text{ as negative}}. \tag{13}$$

There are multiple pedagogical points that instructors can highlight for these measures:

- All these performance measures must be considered before putting a classifier like the LoR model into application.

- The use of PPV and NPV must be made with great care and caution as they are *relative terms*. Both measures depend on the prevalence (overall rate) of malignant tumors. For instance, if most tumors are malignant, we can achieve a high PPV by classifying every tumor as malignant. Alternatively, if malignant tumors are rare, then it will be very difficult to achieve a high PPV. In contrast, sensitivity and specificity do not suffer from this shortcoming and can be used in *absolute terms* as measures of the LoR model's performance.

3.5.2. Performance measures for imbalanced datasets Students have thus far learned how to find an appropriate value for the DB using the performance measures discussed above. These performance measures are often sufficient when data is balanced. However, these measures are insufficient for applications with an inherently imbalanced data, i.e., when the number of 1s (or 0s) in the response variable is considerably lower than its 0s (or 1s).

There are abundant examples such as finding several spam emails (designated with response value 1) among numerous unspammed emails (designated with response value 0) is an example of such imbalanced classification data. In such cases, there are advanced methods to perform classification (e.g., with over or under sampling) and also there exist more appropriate performance measures. While discussing more appropriate classification techniques is outside the scope of this paper, we next explain relevant performance measures for imbalanced data sets. We note that these measures treat precision and recall *asymmetrically* in accordance with their estimated cost (importance).

As a motivating example for the failure of our earlier measures consider the case where one is interested in predicting customers' churn. Then, the "response" rate may be only 1% (which is typical on a monthly basis), the "balanced" accuracy measures all fail as a very accurate prediction is to simply predict all customers as a non-churn. While with a good classification model, a firm can devise effective incentives to reduce churn of valuable customers the previous performance measures cannot bring forward such a good model.

(i) **Decision-boundary-dependent performance measures.** The F_β Scores can be used as a surrogate (or complement) to ROC curve and for a given β is defined as:

$$F_\beta \text{ Score} = (1+\beta^2) \times \frac{(\text{Recall} \times \text{Precision})}{((\beta^2 \times \text{Precision}) + \text{Recall})}, \tag{14}$$

where β is the parameter that determines the weight that the decision maker is inclined to assign to precision versus recall. Note that the F_β Scores is between 0 and 1 and a higher value is better.

Pedagogical points for the F_β Score include:

- F_β Score is mostly used when the distribution of zeros and ones in the response variables is highly *imbalanced*, i.e, the dataset is imbalanced.[8]

[8] More on this subject is the covered in case study sheet.

• Unlike the ROC curve that considers equal weights for both recall and specificity, the F_β Score provides the decision maker with the flexibility of considering different weights for these.

◇ A β value less than 1 assigns a higher weight to the precision than to the recall and a β value greater than 1 assigns higher weight to recall than to the precision.

◇ F1 and F2 scores with β equal to 1 and 2, respectively, are the most commonly-used values of β, and are calculated in our sheet. When $\beta = 1$, the F1 Score is the *harmonic mean* of the precision and recall scores and is used as a metric in scenarios where choosing either of precision or recall score can result in compromise in terms of model giving high FPs and FNs, respectively. F2 Score is used when the weight of recall is twice that of the precision.

• The instructor can compare the F1 Score of DB = 0.5 and DB=0.6 that are 0.80 and 0.89, respectively. The improvement in the F1 score under DB = 0.6 is attributed to the precision value 1.00 rather than 0.80 with DB = 0.5.

• The instructor can finally state that if the data is imbalanced and maximizing the F1 Score is the goal of the classification, DB = 0.6 (or 0.7 or 0.1) provides the highest F1 Score equal to 0.89. DB = 0.6 (or 0.7) coincidentally yields the highest accuracy value of 0.91. Instructors must mention that rarely is the case that one DB concurrently maximizes F1 Score and accuracy performance measures. For instance, DB = 0.8 yields F1 Score and accuracy values of 0.75 and 0.82, respectively.

(ii) **Decision-boundary-independent performance measures.** We introduce here the *lift and decile-wise lift charts*. The values used in these charts are independent of the value of the decision boundaries. These charts are especially informative for *highly imbalanced* data sets. The lift and decile-wise lift charts help determine the success of the classification model.

Both the lift and decile-wise lift charts compare the LoR model's improved classification ability or "lift" with that of a random classification. In order to construct these charts, we sort in a descending manner the probabilities of the classification (in our case based upon the LoR) model.

We first explain the lift chart. For the lift chart the Y-axis is the cumulative number of $1s$. In our example, the Y-axis progressively reaches five—the total number of $1s$ in our dataset. In a dataset with N observations and a proportion of $\bar{y}$ of $1s$, a random classification assigns this proportion to each observation for belonging to class 1. Thus, plotting a straight dashed line with a slope of $\bar{y}$ starting at the point $(0, 0)$ and until the point $(N, N \times \bar{y})$, reflects the expected cumulative proportion of $1s$ observation of a random classification model. In our dataset with 11 observations, the proportion of $1s$ is $\bar{y} = 5/11 = 0.4545$, which is the slope of the random classification line.

The lift chart also includes a curve counting the number of $1s$ observations up to each observation i value; this number is informative as observations appear in descending orders of their probabilities. Obviously, this curve also connects the $(0, 0)$ and $(N, N \times \bar{y})$ points. Moreover, for a good classification model this curve would be above the straight line that corresponds to the

random classifier. The higher this curve is, i.e., the higher is its lift from the line representing the random classifier, the better is the classification model that leads to this curve. In our dataset the first 4 observations with the highest probabilities of the classification of the LoR model, are indeed $1s$, so the curve increases at 45 degrees for these 4 observations and reaches 4 on the Y-axis (see the lift chart in the bottom left corner of Figure 6). Note that the straight line representing the random classification only reaches 4 near the 9th observation. Moreover, the lift curve reaches its top value of 5 at the 7th observation. The fast increase of the curve, i.e., its lift, as well as its long flat tail imply that the LoR classification model is effective (within sample).

We now explain the the Decile-wise lift chart, that investigates the model's classification ability in view of the observations. In order to construct this chart, we divide the sorted observations and data into 10 different groups (deciles), labeled from 1 to 10, on the X-axis. Next, we calculate the mean of the predictive probabilities in each decile, denoted by $\hat{y}_i$. The model's lift within each decile is the ratio between its $\hat{y}_i$ and the expected probability (i.e., a random classification), $\bar{y}_i$ (that is 0.4545 for our dataset). This ratio measures the gain (lift) obtained by the classification model. Mathematically, for each decile we calculate the model's lift as:

$$\textbf{Model's lift}: \frac{\text{The mean of the LoR model's predicted probabilities for each decile}}{\text{The expected response value of 1 for each decile}} = \frac{\hat{y}_i}{\bar{y}}. \quad (15)$$

A similar formula where the numerator is the sum of observations with actual response value of 1 within each decile, $\tilde{y}_i$ provides the actuals' lift:

$$\textbf{Actuals' lift}: \frac{\text{The mean of observations with response value of 1 for each decile}}{\text{The expected response value of 1 for each decile}} = \frac{\tilde{y}_i}{\bar{y}}. \quad (16)$$

The decile-wise lift chart depicts both lifts for each decile.

For our data, we assign one observation for each deciles 1 to 9, and both observations 10 and 11 to decile 10. Then, in the first decile the LoR predicted probability is 0.973 and we divide it by $\bar{y}_i = 0.4545$ to get 2.139 as the model's lift for decile 1. The actuals' lift for decile 1 is $1/0.4545 = 2.2$. Similar calculations are made for the other deciles to get the decile-wise lift chart.

The value of model's lift will reveal how much lift the LoR model has given us in predicting class 1. Higher lift values are indicative of the power of the LoR model in predicting class 1. The similarity between the model and actuals lifts reflects the accuracy of the classification model within each decile in sample.

The important pedagogical moments for the lift and decile-wise charts are:

- The lift/decile charts are used for imbalanced datasets when other standard accuracy models are less informative. When dealing with highly imbalanced dataset (like the churn application) we are mostly concerned with predicted values of the first decile, e.g., the 1% of customers who might churn.

- The predicted bars (orange bars in the decile-wise lift chart) should always show the "*staircase*" pattern. The actual bars (blue bars in the figure) may be in a much more chaotic pattern when the model is not well behaved (in our case because lack of data).
- These charts are used in the context of model selection.

3.6. Model Selection sheet

Now that students are familiar with different classification performance measures, we can discuss model selection. We compare linear regression with logistic regression models on the same data set. This comparison is depicted in Figure 7. [9]

Students learned earlier that linear regression *may* produce output $p(X) < 0$ or $p(X) > 1$ rendering it unsuitable for use as a probabilistic model. Nevertheless, linear regression can be changed into a probabilistic model, as required for classification. Specifically, we change the output of a Linear Regression, $p(X)$ as follows:

1. Round up to zero if $p(X) < 0$
2. Do nothing if $p(X) = [0, 1]$
3. Round down to 1 if $p(X) > 1$

We compare the values after this change against a DB and convert them into zeros and ones similarly to the LoR model. Note that unlike the LoR model where the classification is based upon likelihoods, the linear regression works with the probabilities (obtained after its change).

The objective of this sheet is to help students integrate the material covered in previous sheets. In particular, the pedagogical points for this sheet summarize the points covered in previous sheets as follows:

- Similar to LoR, the optimized coefficients of the linear regression model can be used to create probabilities. For our dataset, the optimized coefficients of the linear regression model are $\hat{\beta}_0 = -3.150$ and $\hat{\beta}_1 = 0.317$. The instructor can remind students that these values have been obtained in the Line-Curve sheet. For example, for observation $x_1 = 11.55$, these coefficients will yield a probability of $\hat{y}_1 = -3.150 + 0.317 \times 11.55 = 0.511$. The instructor must also sensitize students to the fact that the probability for the same observation using LoR is 0.469. This is the point that the instructor can use later to compare the performance of both methods.
- The instructor should point out that for the last observation the prediction of the linear regression is negative $p(9.30)$ or $\hat{y}_{11} = -0.202$ and reiterate that linear regression is not frequently used as a probabilistic model. While not occurring in our dataset, the linear regression model may produce predictions larger than one ($p(X) > 1$) for some observations.

[9] In the Model Selection sheet, we do not include performance measures suited for imbalanced datasets because our small dataset is balanced. We discuss all performance measures in the Case Study sheet where our large dataset is imbalanced.

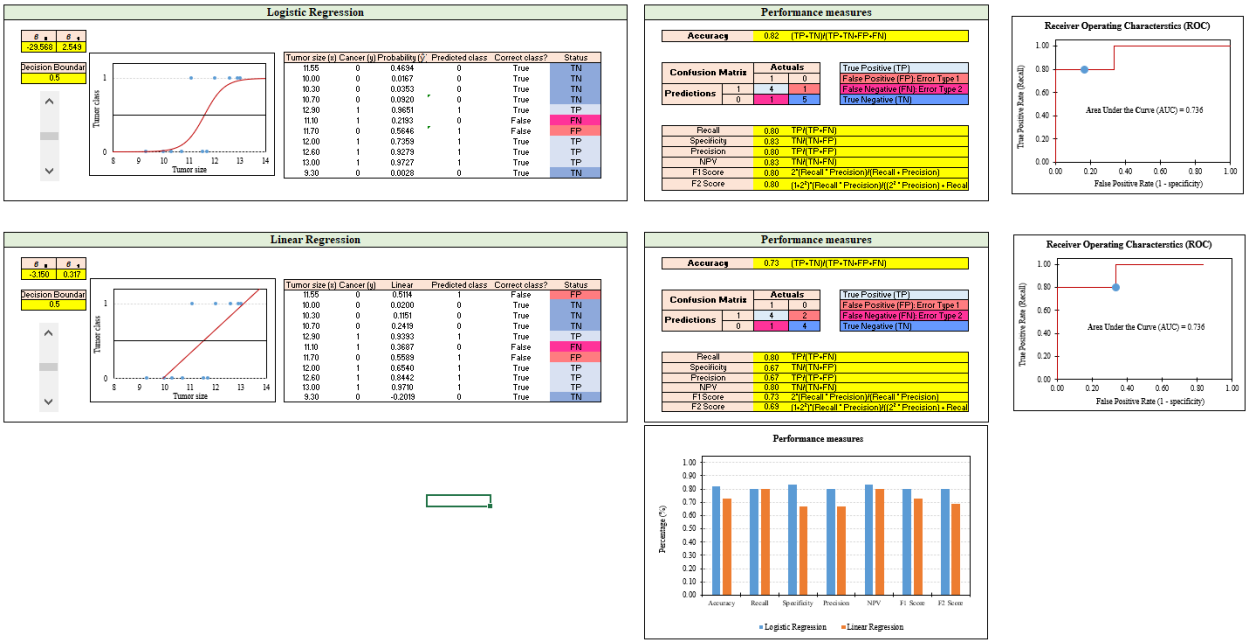


Figure 7 Selection between linear and logistic regression

- Having calculated the probability for each data point, the instructor can now teach students how, using the DB, they can convert these probabilities into zeros and ones that are required for classification. Given the default DB of 0.5, probabilities of greater than and less-than-or-equal to the DB can be classified as 1 and 0, respectively. For instance, the predictions for the first observation $p(11.55)$ is 0.511 and 0.469 in the linear and the logistic regression, respectively. At the default DB of 0.5, these small differences cause the linear regression to erroneously classifies this tumor as malignant, whereas the LoR accurately classifies this tumor as benign.
- The instructor can remind students that working with linear regression is much easier and it does not require working with likelihoods and the MLE for optimizing the likelihood of each observation, but these are easily done in any statistical software.
- Immediately after pointing out the simplicity associated with working with linear regression, the instructor must refer students to the cells where the accuracy of classification for both models (at DB = 0.5) is calculated and highlight that the accuracy of the LoR is 82%, whereas that of linear regression is 73%. Among the many other performance measures at the default DB, the ROC curves indicate that for a True Positive Rate of 0.80, the False Positive Rate of the linear regression model is 0.33, whereas that of logistic regression is 0.17—16% higher Type-1 error due to erroneously classifying benign tumors as malignant ones.
- The instructor can ask students to increase the DB to 0.7 and mention that on the ROC curve, for the FP rate of 0.00, the TP rates of the linear and logistic regressions are 0.60 and 0.80,

respectively—20% higher Type-2 error by the linear regression in that malignant tumors are erroneously classified as benign ones, which can have dire life consequences for patients with malignant tumors.

• One counter-intuitive point for students is that the FP rate on the x-axis of the ROC curve associated with the linear regression is curtailed at 0.80 and never reaches 1.00. This phenomenon occurs when the DB is set to zero (DB = 0) in which case all observations with $p(x) \geq 0$ are classified as malignant. Since the linear regression model has an observation with negative probability $p(9.30) < 0 = -0.202$, that observation is never recognized as malignant. If one is interested in remedying this issue and having an uncurtailed ROC curve that extends to TP rate of 1.00 for the linear regression model, the DB must be set to a negative number, slightly smaller than the smallest negative predictions across all observations.

• The instructor should point that the values of Area Under the Curve (AUC) for the ROC curve is independent of the DB. Moreover, these values for both models are identical (if we assume the linear FP rate could be extended up to 1.0). Higher AUC values indicate that the classifier better distinguishes between benign and malignant tumors. However, despite identical AUC values for both models, they yield different performances under different DBs. For instance, the LoR model achieves a TP rate of 0.80 and FP rate of 0.00 at DB = 0.7, whereas the linear regression model achieves these performances at DB = 0.6. This comparison demonstrates that, if AUC values are the same for both models, similar trade-offs can be obtained via different DBs. So, the choice of DB is of paramount importance.

• Remind students that no model selection should be exclusively done within the *In-Sample* data. Once we fixed the DB and optimized the parameters of our models using the In-Sample data, we should evaluate its performance on the Out-of-Sample data and if we are not satisfied with this performance, we should go readjust the DB or develop and alternative model. We further explain this concept in the next sheet where we also discuss over-fitting.

3.7. Cross Validation sheet

This sheet focuses on cross validation using the LoR model. We optimized the coefficients of the LoR function, β_0 and β_1, using a single data set (known as **Training** or *In-Sample* data set) as provided in sheet "**Data**". Students must be reminded that these optimized coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ are later used to predict whether the tumor of a **new** patient is benign or malignant. A correct classification will save lives of and mental stress for patients. Therefore, as for any other prediction model, one must scrutinize the performance of the LoR model on *Out-of-Sample* data (known as **Test sets**). In our context such data represents potential future patients. The process of ascertaining the robustness of any statistical model and its applicability to any related data sets, is called **Cross**

Validation.[10] In this Cross-Validation sheet, we provide students with the information of five new patients with tumor and their statuses. Students can first observe how these tumors are classified given the $\hat{\beta}_0$ and $\hat{\beta}_1$ obtained from the training set for a given DB. This sheet allows students to vary the DB and view its impact on performance measures of the test and training sets.

If the accuracy of the LoR model is much higher in the training set than in the test set, the instructor can point out the issue of ***over-fitting*** *of the LoR model with respect to the training set.* Over-fitting occurs when the performance of the LoR method on the training set is substantially superior to those of the test set. Over-fitting, indicates that the LoR model has not been properly calibrated for actual classification purposes. Instructors can use the DB of 0.5, where the accuracy of the LoR model on the training and test sets is 0.82 and 0.80 (closest performance), respectively. Then, use a DB of 0.8, where these accuracies are 0.82 and 0.60 (sharply different performance), respectively.

This sheet, depicted at Figure 8, allows students to dynamically measure the impact of the DB on performance measures and investigate the issue of over-fitting. Once students were sensitized to the out-of-sample performance of the LoR model, the instructor can mention that there will be a separate lecture on the subject of cross validation and re-calibration of LoR coefficients (if such a lecture is planned).

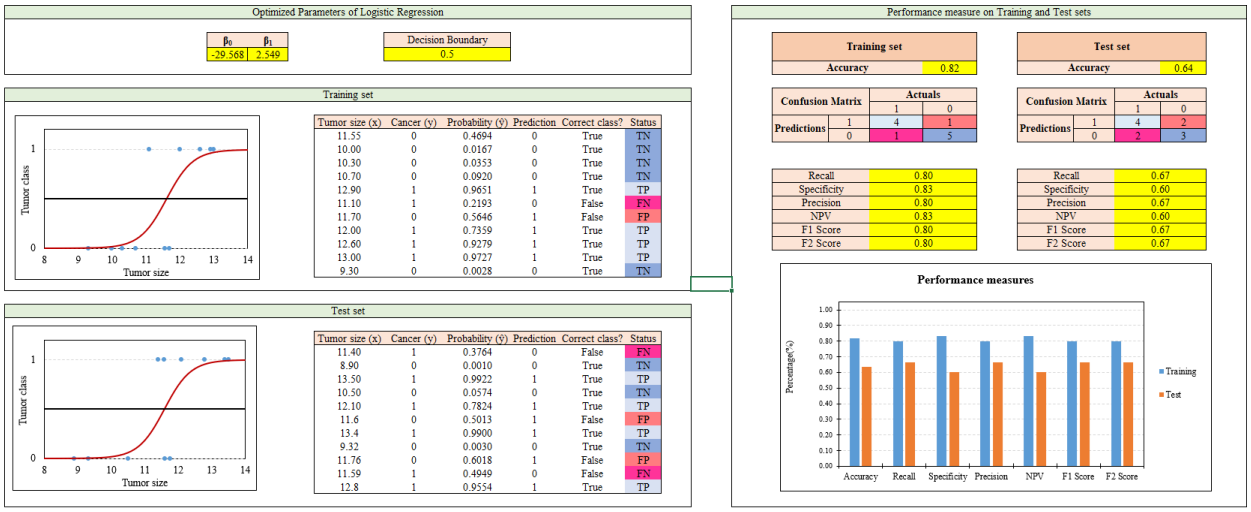


Figure 8 Cross Validation

Remark. *The purpose of our classification is to calibrate our LoR model in such a way that it performs effectively on the test (out-of-sample) set. In the Performance Measures sheet, we used the entire data to calculate performance measures. Now that students are familiar with the concept of the training and test sets, we continue our explanation based on the fact that they are applied to the training set. As such, all the related optimization of the LoR model's coefficients are performed on the test set.*

[10] There are many ways to cross-validate a model. We do not review these.

3.8. Case Study sheet

Our entire discussion on LoR has thus far been focused on a *single-predictor* logistic regression. Once students learned how to use the excel solver (in MLE sheet) to optimize the coefficients of a single-predictor LoR model, they are well-equipped to optimize coefficients of a logistic regression with *multiple predictors*. The formula for the LoR model with p predictors is:

$$p(X) = \frac{1}{1 + e^{-\left(\beta_0 + \sum_{i=1}^{p} \beta_i X_i\right)}}, \tag{17}$$

where β_i is the coefficient assigned to predictor X_i.

We prepared the Case Study sheet, depicted in Figure 9, to support the LoR model with up to 10 predictors and up to 10,000 observations. The dataset should be inserted in cells E13:O1012.

We demonstrate the usage of this sheet using a case study related to calculating the *probability of a customer's credit default* as a function of three predictors: customers being students or not (X_1), the amount of balance on their credit card (X_2), and their income (X_3). We use the publicly available dataset of "default" introduced in James et al. (2014). This dataset consists of $10,000$ rows and four columns. Response values designated as zeros and ones have been included under the Y column. Predictors X_1 (categorical), X_2 (nominal), and X_3 (nominal) constitute other columns of this dataset. Note that the amount of income has been divided by 1,000. While analyzing this dataset, we use all the classification performance measures that students have learned in previous sheets.

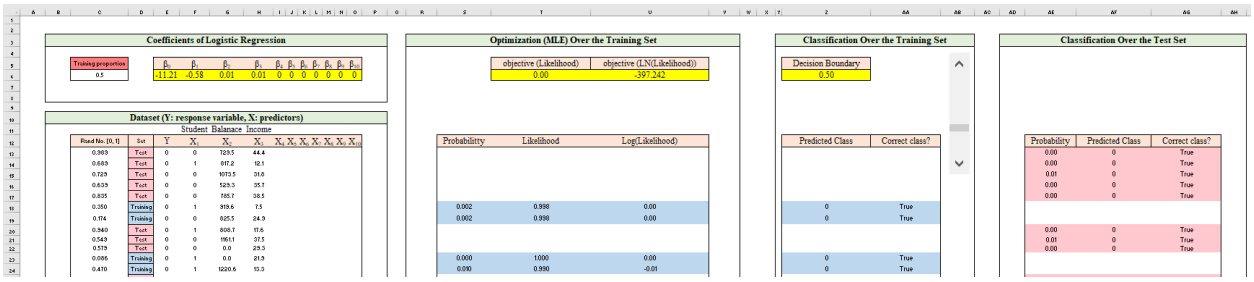


Figure 9 Case study: Logistic Regression with multiple predictors

The optimized coefficients of the LoR and its Maximum Likelihood are printed in Cells E6:O6 and U6, respectively. To obtain these coefficients, the instructor can go to **Data** $\rightarrow$ **Solver** in which case the panel in Figure 10 will pop out. This figure shows that the objective function is of maximization type and the cells that coefficients and objective function values are printed. To find optimized coefficients, students only require to press "Solve" button. The solver assigns value of zero to those predictors for which no data entry has been made.

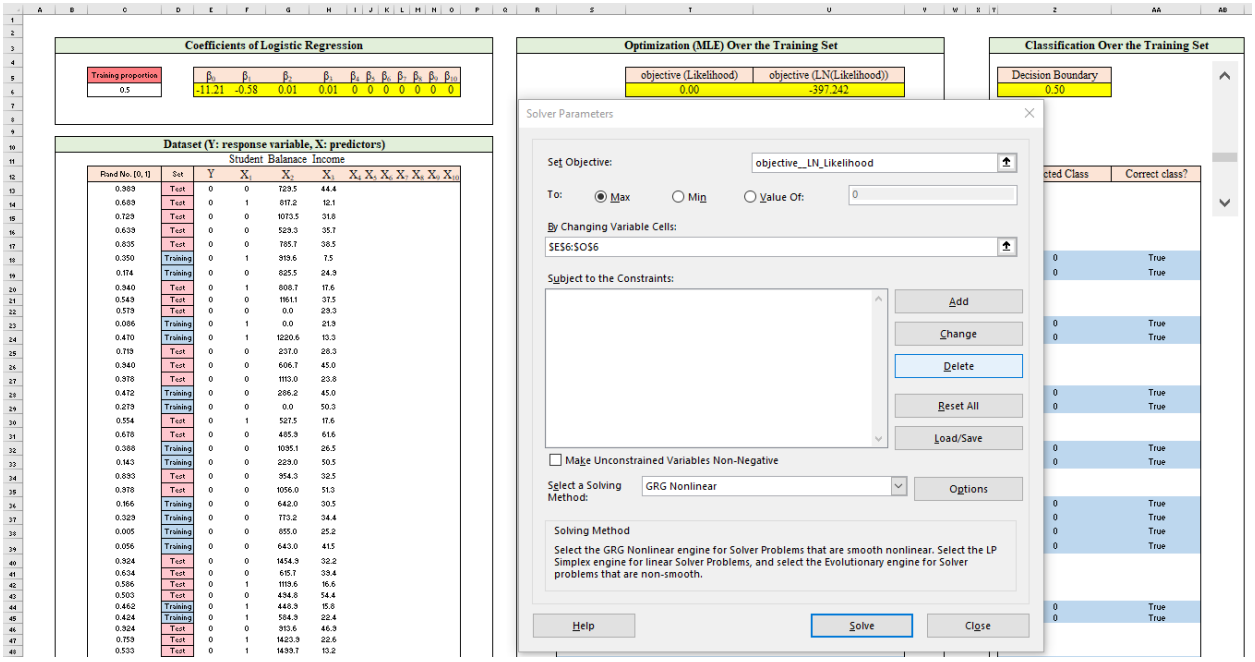


Figure 10 Finding coefficients of the logistic regression with multiple predictors using Excel solver

There are several detailed pedagogical points about this sheet that instructors can emphasize. Instructors can use their discretion to overemphasize or de-emphasize some of these pedagogical points based on the nature of the course and learning objectives of the lecture. Below, we provide an excerpt of potential learning objectives for this sheet.

- Students have already learned (in the Cross-Validation sheet) that the LoR model is optimized using the training set and it is evaluated on the training (in-sample) and test (out-of-sample) sets—with placing more significant emphasis on the accuracy rate of the test set. Thus, the choice of training data set impacts the coefficients of the LoR model and hence its accuracy on the test set. The case study sheet allows instructors and students to flexibly select the training data set by determining the value of the "Training Proportion" in cell C5 (see Figure 9). The instructor can set the Training Proportion to (e.g.) 0.75 in which case 75% ($\approx$ 7,500 out of 10,000) of data points are randomly selected as training set. The instructor must remind students that a random number in the range of [0, 1] has been generated and preassigned to each data point. If the assigned random number is less than or equal to the value of the "Training" proportion, that data point is highlighted in blue in Column D and considered as part of the training set; other data points are highlighted in pink and are a part of the test set. This interactive usage of different colors for the training and test sets allow students to understand better the impact of "Training Proportion" on the coefficients of the LoR model.
- The instructor must elucidate the fact that the optimized LoR coefficients can be used to determine the accuracy of classification on both the training and test sets. The value of the **training**

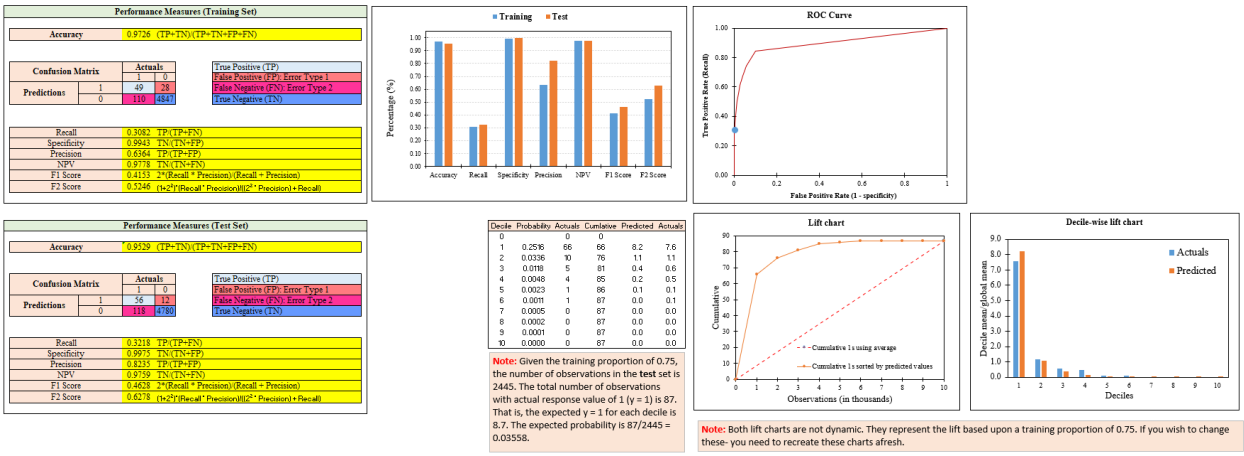
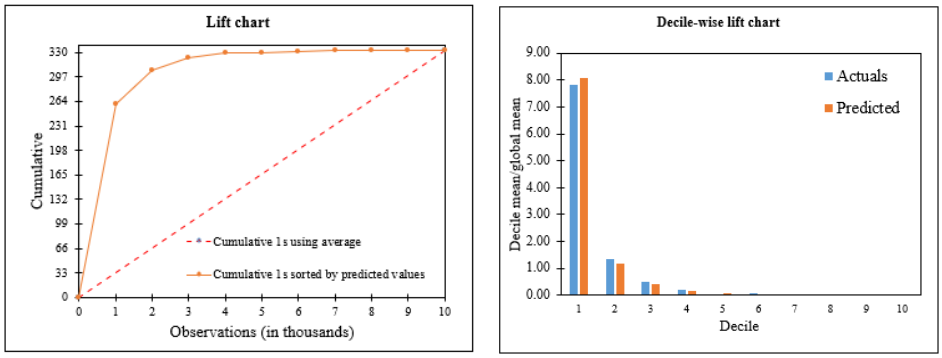


Figure 11 Performance measures of the LoR model at training proportion of 0.75 and DB value of 0.5.

proportion and **DB** should yield similar accuracy rates in the training and test sets, i.e., avoiding over-fitting. The performance of the LoR model given the value of training proportion and a DB can be evaluated on the performance measures that were discussed in previous sections (Figure 11). This sheet provides the instructor with a sufficient leeway to cover all performance measures discussed in Sections 3.5 and 3.7. For instance, the instructor can set the Training Proportion to 0.5 and re-optimize the coefficients of the LoR model. Then, for most values of the DB, the performance of the LoR model on the training and test sets are close.

- Having optimized the coefficients of the LoR model at the training proportion of 0.75 and set DB to its default value at 0.5, the instructor can mention that the accuracy of the LoR model (on the training set) is 0.9734 and ask students whether this is good level of accuracy or not. Most students will likely respond that this level of accuracy is really high for a classifier. Once such a response was received from any student, the instructor can immediately point out that this dataset contains only 333 defaulting customers with $Y = 1$ and $9,667$ non-defaulting customers with $Y = 0$. Thus, for this *imbalanced* dataset (that contains many more zeros than ones as the response variable), an accuracy of 96.67% can be easily achieved by predicting that all observations as non-defaulting ones ($0.9667 = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} = \frac{0+9,667}{0+9,667+0+333}$). Thus, *the judgement regarding the classification model's accuracy must be made in the context of the given dataset.*
- The instructor should refer students to the decile/lift charts. These two charts must be used together to demonstrate how well the LoR classifier has been able to predict response values. We expect 33.3 observations with responses $Y = 1$ (diagonal red dashed line in Figure 12(a)) for each decile if we use the average model that has no predictive power. The use of the LoR model (at training proportion equal to 1) helps correctly classify 261 out of 333 defaulting customers (78.37%) in the first decile. This shows that the LoR model gives lift to the accuracy of the classification by the factor of 8 for the first decile (Figure 12(b)).



(a) Decile chart (b) Decile-wise lift chart

Figure 12 DB-independent decile/lift charts

- Once previous pedagogical points were covered, the instructor can reemphasize that no performance measure must be used in vacuum for ascertaining the performance effectiveness of a classifier. It is thus imperative that various performance measures be scrutinized concurrently and in the context of the data.

Finally, as in-class or take-home exercise, the instructor can ask students to use the linear regression with multiple predictors to perform the classification on the dataset. Furthermore, the instructor can ask students to capture the interaction among predictors (i.e., $X_4 = X_1 \cdot X_2$, $X_5 = X_1 \cdot X_3$, $X_6 = X_2 \cdot X_3$, and $X_7 = X_1 \cdot X_2 \cdot X_3$) on the performance of the LoR model.

4. Conclusion

In this paper, we developed a comprehensive and self contained interactive spreadsheet model that facilitates the teaching of logistic regression and classification to students in science, analytics, statistics, business, and engineering. Interactive features of this spreadsheet model allow students to play around with different parameters of the logistic regression models and by so to study the notion of maximum likelihood and the impact of different classifications (e.g., by changing the DB) on different performance measures. These interactive features allow students to develop a strong understanding of classifiers' performance measures and enable them to determine the best decision boundary that leads to the most desired performance of the logistic regression model (or other classification models). We discussed two sets of performance measures for balanced and imbalanced datasets. We further provided a multivariate logistic regression case study that helps instructors to implement such classification on new data sets. We use this case study to analyze data illustrate that performance measures of a classifier should be interpreted within the context of the application. Overall, our spreadsheet model will enhance students' comprehension of maximum likelihood estimation, logistic regression, classification techniques, and model selection.

Appendix

We review the MLE Explained sheet that is used to optimize coefficients of the LoR model. This sheet can be taught after the likelihood sheet discussed in Section 3.3. We assume students are familiar with basic optimization concepts.

MLE Explained sheet

In this sheet we detail the algorithmic procedure, MLE, that is used to find optimal values for the coefficients of the LoR function: $\hat{\beta}_0$ and $\hat{\beta}_1$. (After students are familiarized with calculating probability and likelihood in the "Likelihood" sheet. Sheet "**MLE Explained**" integrates all previous learnings and provides a procedural framework that exposes students to the steps involved in calculating MLE. Students previously learned that Equation (5) has a logit that is linear in X:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X.$$

We first calculate the logit of this function that is obtainable by any combination value of β_0, β_1, and X. We call this value, logit $(\beta_0 + \beta_1 X)$. Second, we exponentiate the logit function given any combination value of β_0, β_1, and X to define and obtain $d := e^{-(\beta_0+\beta_1 X)}$. Now, we calculate the *probability* of each observation i, $\hat{y}_i$ using formula $\frac{1}{(1+d)}$, i.e., $\hat{y}_i = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}}$. We multiply these probabilities to calculate the likelihood estimate that we then maximize as

$$\text{Maximum Likelihood}: \;\; L(\beta_0, \beta_1) = \prod_{i \in \mathcal{I}} \hat{y}_i^{(y_i)} \times (1-\hat{y}_i)^{(1-y_i)}.$$

Note that this is the same formula that we have already presented in Equation (8). Remind students that directly maximizing the above equation is difficult due to the *multiplicative* relationship between the likelihood of different data points. Furthermore, we already observed $L(\beta_0, \beta_1)$ in the dynamic stacked bar chart in sheet "Likelihoods".

We convert this multiplicative relationship into an *additive* relationship that is amenable to optimization (e.g., using Excel Solver), by taking logarithm (ln) to get:

$$\text{Sum of Log Likelihoods}: \;\; \ell(\beta_0, \beta_1) = \sum_{i \in \mathcal{I}} y_i \ln(\hat{y}_i) + (1-y_i)\ln(1-\hat{y}_i). \tag{18}$$

This additive function is more amenable to optimization than the multiplicative one and students can optimize the coefficients of β_0 and β_1 using the Excel solver. Specifically, students can go to Data and click on Solver in which case they will see that cell H3 (in Figure 13) is designated as the objective function in which the value of the additive objective function is populated. The range of the objective function value of the additive model is $(-\infty, 0)$—values close to zero indicates a better fit of the nonlinear curve to data points. Cell B3:C3 (in Figure 13) have been designated

for the coefficients of LoR method. Once the coefficients of the additive model was optimized, one can calculate likelihoods of observation and multiply them together and obtain the objective function value of the first multiplicative objective function value, which appears in cell G3.

β_0	β_1
-29.567	2.549

objective (Likelihood)	objective (LN(Likelihood))
0.028	-3.579

		Blackbox		Probability	Likelihood	LN(Likelihood)
Tumor size (x)	Cancer (y)	Logit ($\beta_0 + \beta_1 x$)	d =$e^{(-1 * \text{logit})}$	$\hat{y}$ = 1/(1+d)	$\hat{y}$^(y)*(1-$\hat{y}$)^(1-y)	y*LN($\hat{y}$)+(1-y)*LN(1-$\hat{y}$))
11.55	0	-0.1240	1.1320	0.469	0.5310	-0.6331
10.00	0	-4.0752	58.8616	0.017	0.9833	-0.0168
10.30	0	-3.3104	27.3972	0.035	0.9648	-0.0358
10.70	0	-2.2908	9.8826	0.092	0.9081	-0.0964
12.90	1	3.3174	0.0362	0.965	0.9650	-0.0356
11.10	1	-1.2711	3.5648	0.219	0.2191	-1.5184
11.70	0	0.2584	0.7723	0.564	0.4358	-0.8307
12.00	1	1.0231	0.3595	0.736	0.7356	-0.3071
12.60	1	2.5526	0.0779	0.928	0.9278	-0.0750
13.00	1	3.5723	0.0281	0.973	0.9727	-0.0277
9.30	0	-5.8596	350.5860	0.003	0.9972	-0.0028

Step 1: Sigmiod value $\hat{y} = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$

Step 1.1: $(\beta_0 + \beta_1 x)$
Step 1.2: $e^{-(\beta_0+\beta_1 x)}$
Step 1.3: $\frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$

Step 2: Likelihood
y * LN($\hat{y}$) + (1-y) *LN(1-$\hat{y}$)

Figure 13 Maximum Likelihood Estimation (MLE) Method

References

Boutilier, J. J., T. C. Y. Chan. 2021. Introducing and integrating machine learning in an operations research curriculum: An application-driven course. *INFORMS Transactions on Education* **0**(0) null. doi:10.1287/ited.2021.0256. URL https://pubsonline.informs.org/doi/abs/10.1287/ited.2021.0256.

Brusco, M. 2021. Logistic regression via excel spreadsheets: Mechanics, model selection, and relative predictor importance. *INFORMS Transactions on Education* **0**(1-11) null. doi:10.1287/ited.2021.0263. URL https://doi.org/10.1287/ited.2021.0263.

Dan, T., P. Marcotte. 2019. Competitive facility location with selfish users and queues. *Operations Research* **67**(2) 479–497.

Erkut, E., A. Ingolfsson. 2000. Let's put the squares in least-squares. *INFORMS Transactions on Education* **1**(1) 47–50.

Huggins, E., M. Bailey, I. Guardiola. 2020. Case article—converting NFL point spreads into probabilities: A case study for teaching business analytics. *INFORMS Transactions on Education* **21**(1) 57–60.

James, G., D. Witten, T. Hastie, R. Tibshirani. 2014. *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated.

Jiang, G., L. J. Hong, B. L. Nelson. 2020. Online risk monitoring using offline simulation. *INFORMS Journal on Computing* **32**(2) 356–375.

Kim, D., N. Kim, J. Cho, H. Shin. 2019. Optimizing the multistage university admission decision process. *INFORMS Journal on Applied Analytics* **49**(6) 422–429.

Kopcso, D., Dessislava P. 2018. Case article—business value in integrating predictive and prescriptive analytics models. *INFORMS Transactions on Education* **19**(1) 36–42.

Liberatore, M., R. Nydick, C. Daskalakis, E. Kunkel, J. Cocroft, R. Myers. 2009. Helping men decide about scheduling a prostate cancer screening exam. *INFORMS Journal on Applied Analytics* **39**(3) 209–217.