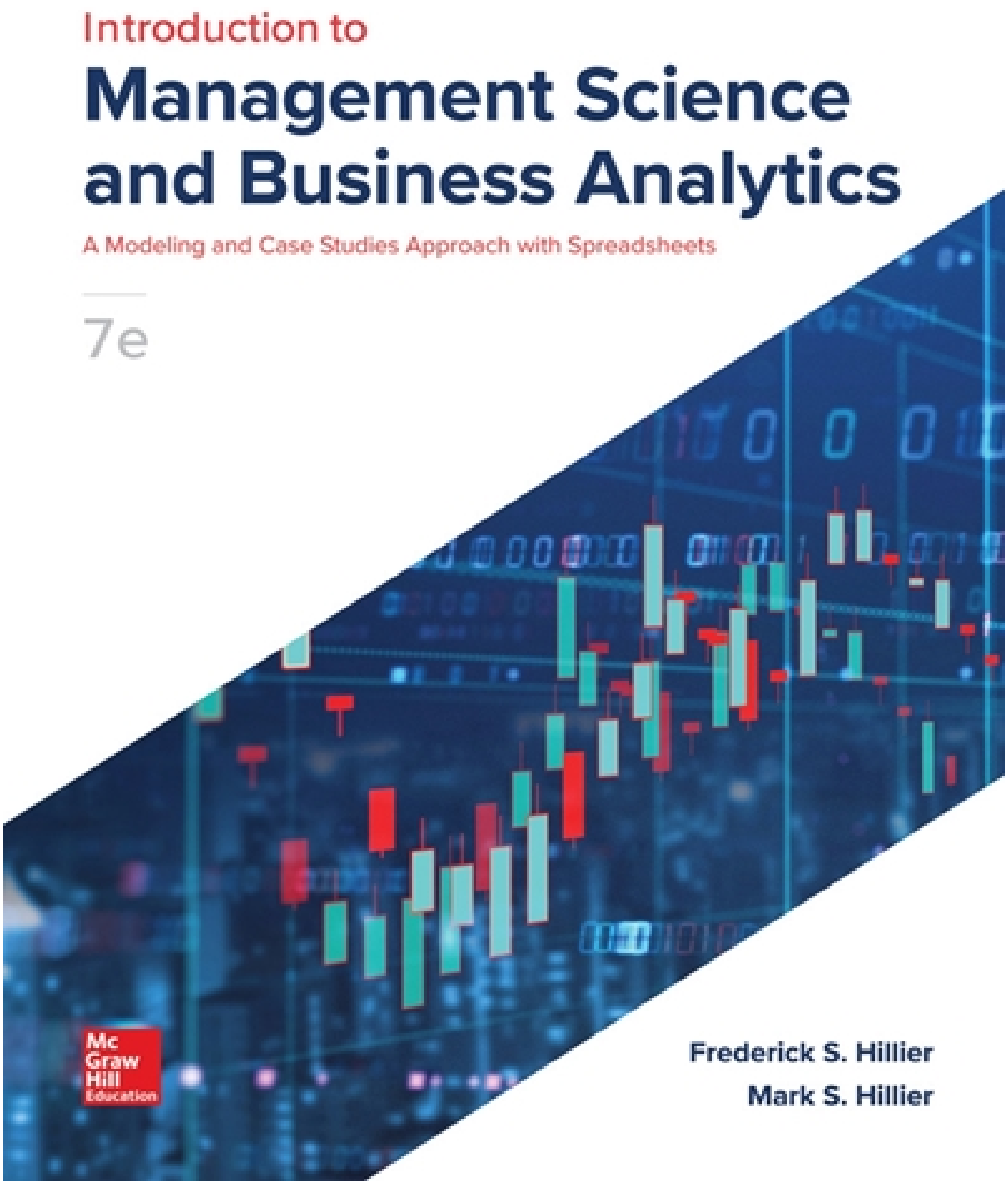
# Solutions for Introduction to Management Science and Business Analytics 7th Edition by Hillier


Introduction to
Management Science and Business Analytics
A Modeling and Case Studies Approach with Spreadsheets
7e
Mc Graw Hill Education
Frederick S. Hillier
Mark S. Hillier

# Solutions

# CHAPTER 2
# OVERVIEW OF THE ANALYSIS PROCESS
# SOLUTION TO SOLVED PROBLEM

## 2.S1 Category D Loan Applicants at First Bank

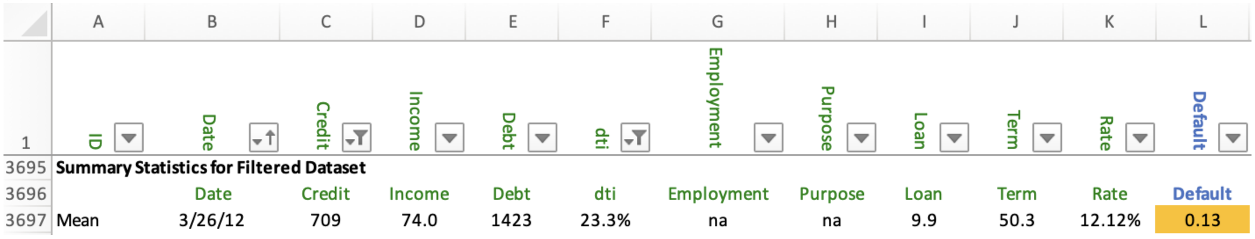
*In the First Bank case study considered in this chapter, applicants for unsecured loans are classified as category D if they have a high enough expected default rate that the loan should be denied. The CEO and the VP of the loan department have determined that applicants that are estimated to have a greater than 25 percent chance of default should fall into category D and be denied a loan. Use the dataset on the* Clean Data *tab of the* First Bank Data *spreadsheet available at* **www.mhhe.com/Hillier7e** *to answer each of the following questions.*

*a. First Bank is considering requiring Credit ≥ 600 and dti ≤ 36 percent to approve a loan (and avoid category D). Filter the dataset and use the AGGREGATE function to determine what the average default rate would have been for historical customers that met each of the following criteria.*

*(i) Credit ≥ 600 and dti ≤ 36 percent*
*(ii) Credit < 600*
*(iii) dti > 36 percent*

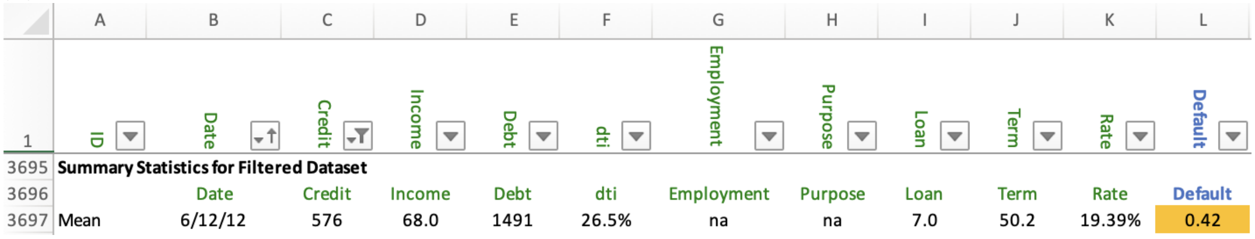
On the Clean Data tab, the formula =AGGREGATE(1, 5, L2:L3685) in cell L3697 calculates the mean value of Default for the filtered dataset. Applying a filter to Credit and dti for each of the criteria *(i)*, *(ii)*, and *(iii)* yields the following average default rates.

*(i)* Mean value of Default = 0.13



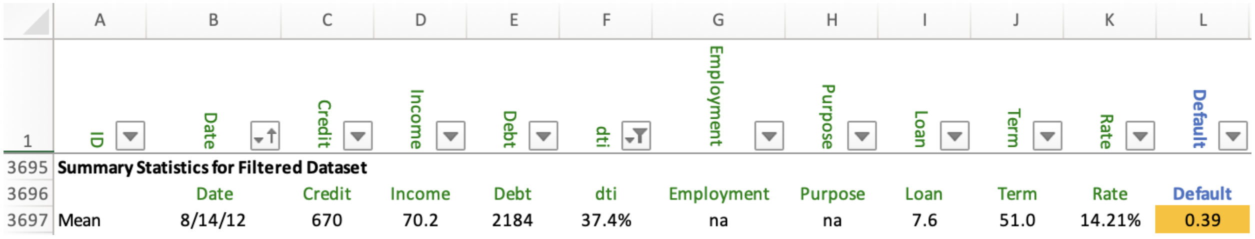
| | A | B | C | D | E | F | G | H | I | J | K | L |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ID | Date | Credit | Income | Debt | dti | Employment | Purpose | Loan | Term | Rate | Default |
| 3695 | **Summary Statistics for Filtered Dataset** | | | | | | | | | | | |
| 3696 | | Date | Credit | Income | Debt | dti | Employment | Purpose | Loan | Term | Rate | Default |
| 3697 | Mean | 3/26/12 | 709 | 74.0 | 1423 | 23.3% | na | na | 9.9 | 50.3 | 12.12% | 0.13 |

*(ii)* Mean value of Default = 0.42



| | A | B | C | D | E | F | G | H | I | J | K | L |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ID | Date | Credit | Income | Debt | dti | Employment | Purpose | Loan | Term | Rate | Default |
| 3695 | **Summary Statistics for Filtered Dataset** | | | | | | | | | | | |
| 3696 | | Date | Credit | Income | Debt | dti | Employment | Purpose | Loan | Term | Rate | Default |
| 3697 | Mean | 6/12/12 | 576 | 68.0 | 1491 | 26.5% | na | na | 7.0 | 50.2 | 19.39% | 0.42 |

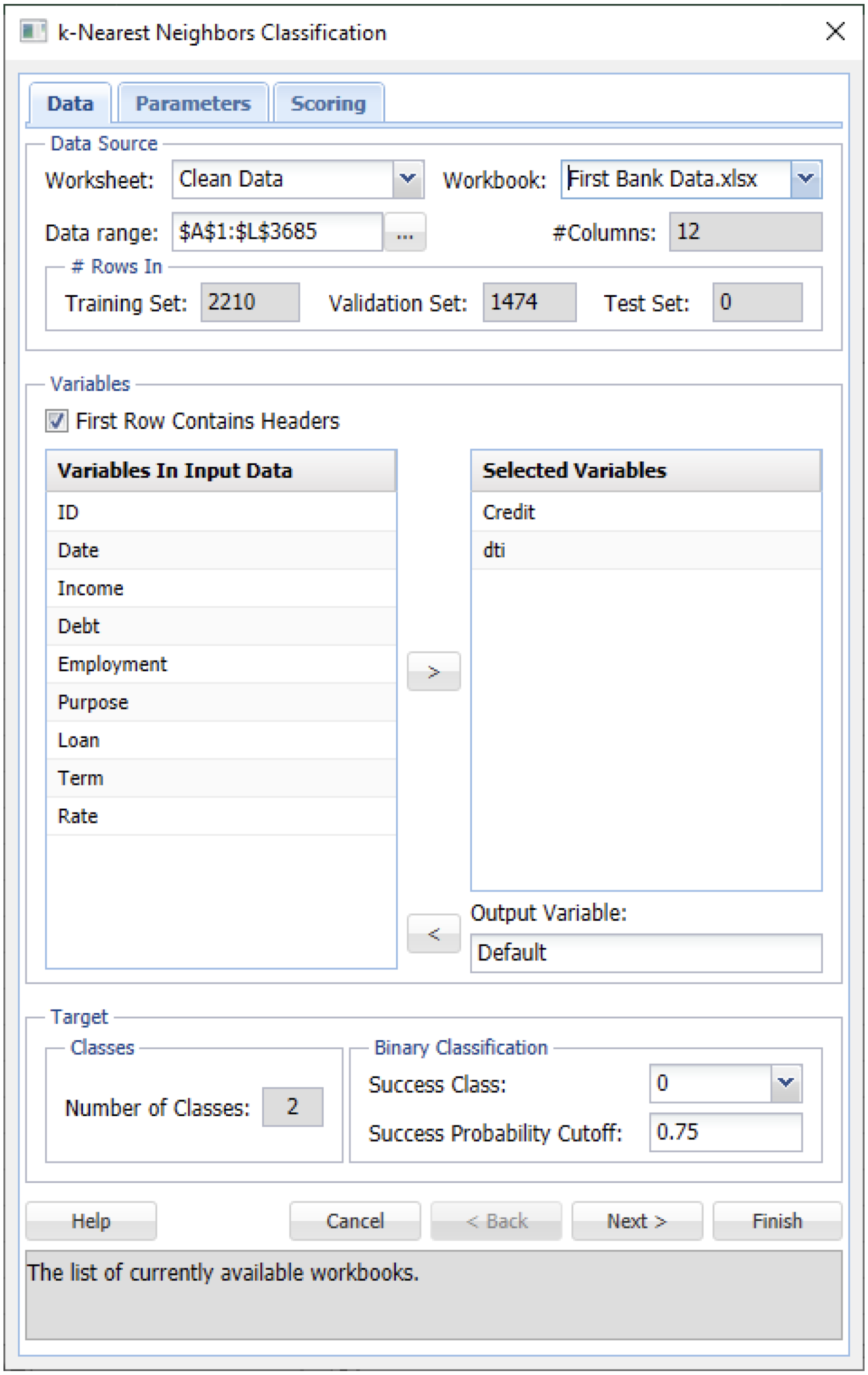
*(iii)* Mean value of Default = 0.39



| | A | B | C | D | E | F | G | H | I | J | K | L |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ID | Date | Credit | Income | Debt | dti | Employment | Purpose | Loan | Term | Rate | Default |
| 3695 | **Summary Statistics for Filtered Dataset** | | | | | | | | | | | |
| 3696 | | Date | Credit | Income | Debt | dti | Employment | Purpose | Loan | Term | Rate | Default |
| 3697 | Mean | 8/14/12 | 670 | 70.2 | 2184 | 37.4% | na | na | 7.6 | 51.0 | 14.21% | 0.39 |

b. *Partition the historical records into a training partition (60 percent of the records) and a validation partition (the remaining 40 percent of the records). Use Credit and dti as predictor variables with the data standardized. Determine the accuracy, sensitivity, and specificity for the KNN model with k = 10 (based on the historical data in the training partition), when it is used to classify applicants as to whether they are likely to default (defined as greater than a 25 percent chance).*

Choose *k-Nearest Neighbors* from the *Classify* menu of the *Data Mining* tab of *Analytic Solver*. Specify the predictor variables as *Credit* and *dti*, the output variable as *Default*, the *Success Class* as 0, and the *Success Probability Cutoff* as 0.75, as shown in the dialog box below.



k-Nearest Neighbors Classification

Data Source

Worksheet: Clean Data

Workbook: First Bank Data.xlsx

Data range: $A$1:$L$3685

#Columns: 12

# Rows In

Training Set: 2210

Validation Set: 1474

Test Set: 0

Variables

☑ First Row Contains Headers

| Variables In Input Data |
|---|
| ID |
| Date |
| Income |
| Debt |
| Employment |
| Purpose |
| Loan |
| Term |
| Rate |

| Selected Variables |
|---|
| Credit |
| dti |

Output Variable: Default

Target

Classes

Number of Classes: 2

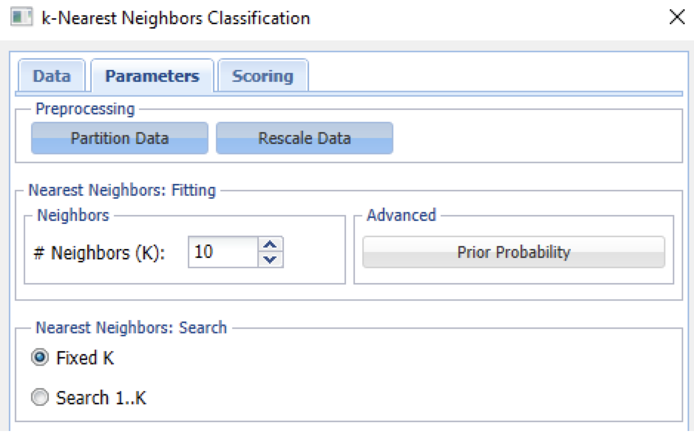
Binary Classification

Success Class: 0

Success Probability Cutoff: 0.75

The list of currently available workbooks.

On the *Parameters* tab of the dialog box, choose *K* = 10 and *Fixed K* as shown below.



k-Nearest Neighbors Classification

Data | Parameters | Scoring

Preprocessing: Partition Data | Rescale Data

Nearest Neighbors: Fitting

Neighbors: # Neighbors (K): 10

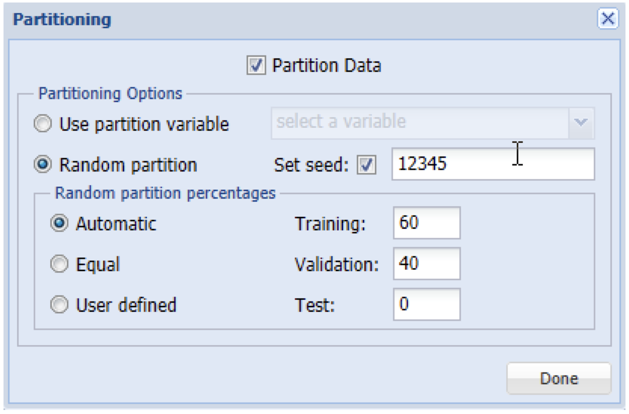
Advanced: Prior Probability

Nearest Neighbors: Search

● Fixed K

○ Search 1..K

Also choose *Partition Data* and *Rescale Data* and fill in the resulting dialog boxes as shown below to partition the dataset and rescale the data using *Standardization*.



Partitioning

☑ Partition Data

Partitioning Options

○ Use partition variable — select a variable

● Random partition — Set seed: ☑ 12345

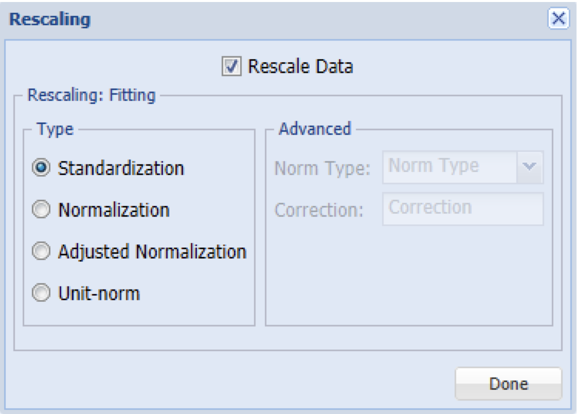
Random partition percentages

● Automatic — Training: 60

○ Equal — Validation: 40

○ User defined — Test: 0

Done



Rescaling

☑ Rescale Data

Rescaling: Fitting

Type

● Standardization

○ Normalization

○ Adjusted Normalization

○ Unit-norm

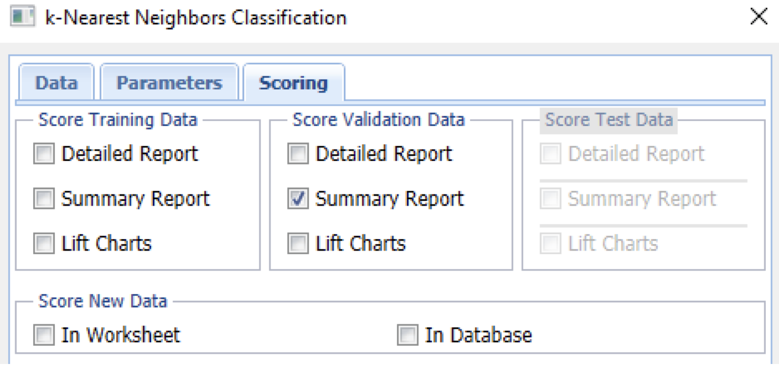
Advanced

Norm Type: Norm Type

Correction: Correction

Done

On the *Scoring* tab of the dialog box (shown below), choose *Summary Report* for the validation data to show the results of applying the KNN algorithm to the validation data.



k-Nearest Neighbors Classification

Data | Parameters | Scoring

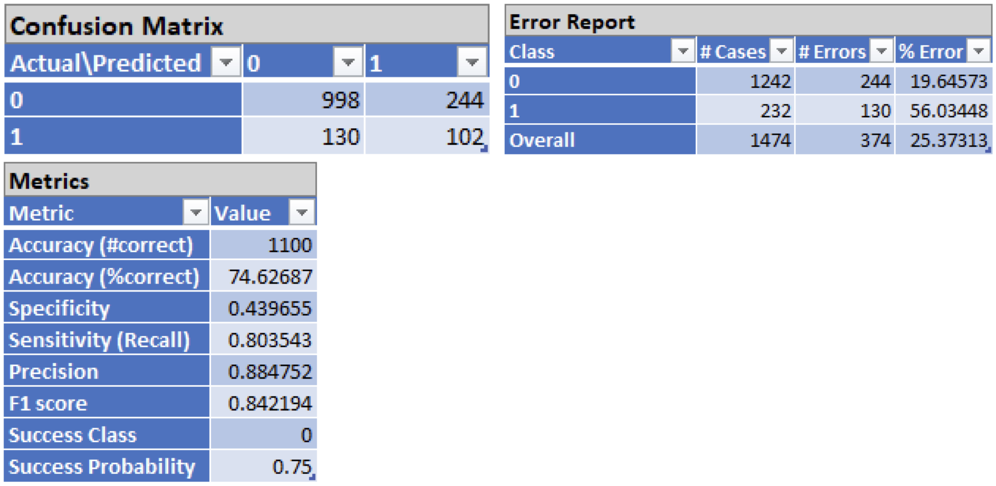
Score Training Data: □ Detailed Report □ Summary Report □ Lift Charts

Score Validation Data: □ Detailed Report ☑ Summary Report □ Lift Charts

Score Test Data: □ Detailed Report □ Summary Report □ Lift Charts

Score New Data: □ In Worksheet □ In Database

This leads to the summary of results shown below. As indicated in the *Metrics* table, the accuracy is 74.6 percent, the specificity is 44.0 percent, and the sensitivity is 0.80 percent.



**Confusion Matrix**

| Actual\Predicted | 0 | 1 |
|---|---|---|
| 0 | 998 | 244 |
| 1 | 130 | 102 |

**Error Report**

| Class | # Cases | # Errors | % Error |
|---|---|---|---|
| 0 | 1242 | 244 | 19.64573 |
| 1 | 232 | 130 | 56.03448 |
| Overall | 1474 | 374 | 25.37313 |

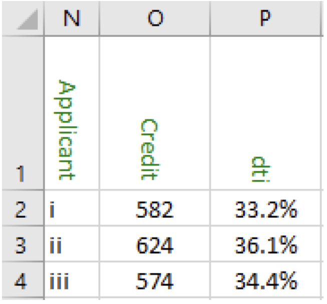
**Metrics**

| Metric | Value |
|---|---|
| Accuracy (#correct) | 1100 |
| Accuracy (%correct) | 74.62687 |
| Specificity | 0.439655 |
| Sensitivity (Recall) | 0.803543 |
| Precision | 0.884752 |
| F1 score | 0.842194 |
| Success Class | 0 |
| Success Probability | 0.75 |

c. *Using Credit and dti as predictor variables and the data standardized, apply the KNN algorithm to this problem with k = 10 to the entire dataset (unpartitioned). Classify each of the following applicants as likely to default (defined as greater than a 25 percent chance) and indicate each applicant's estimated probability of default.*

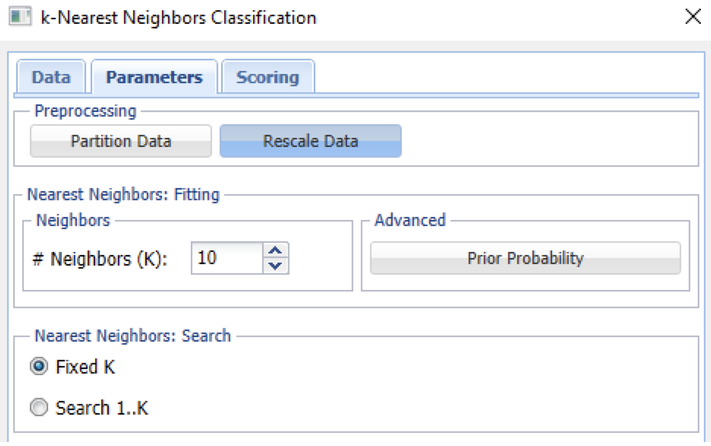
*(i) Credit = 582 and dti = 33.2 percent.*
*(ii) Credit = 624 and dti = 36.1 percent.*
*(iii) Credit = 574 and dti = 34.4 percent.*

Enter data for the new applicants on the worksheet as shown below.



|   | N | O | P |
|---|---|---|---|
| 1 | Applicant | Credit | dti |
| 2 | i | 582 | 33.2% |
| 3 | ii | 624 | 36.1% |
| 4 | iii | 574 | 34.4% |

Choose *k-Nearest Neighbors* from the *Classify* menu of the *Data Mining* tab of *Analytic Solver*. Fill in the *Data* tab of the dialog box exactly the same as in part *b*. On the *Parameters* tab of the dialog box, choose *K* = 10 and *Fixed K* as in part *b*. Turn off partitioning to use the entire dataset. Rescale the data using *Standardization*.



k-Nearest Neighbors Classification

Preprocessing

Nearest Neighbors: Fitting

Neighbors

# Neighbors (K): 10

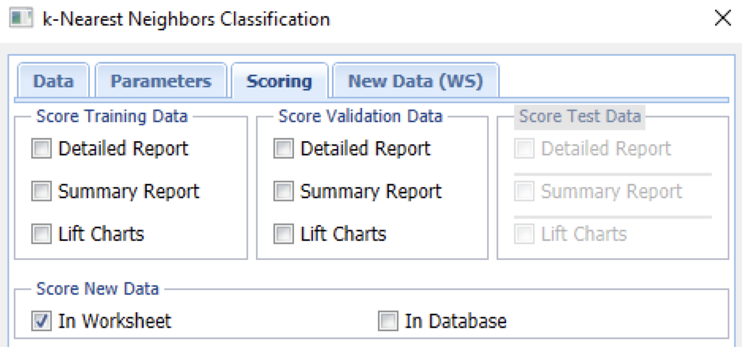
Advanced

Nearest Neighbors: Search

● Fixed K

○ Search 1..K

On the *Scoring* tab, check to box to score new data in the worksheet.



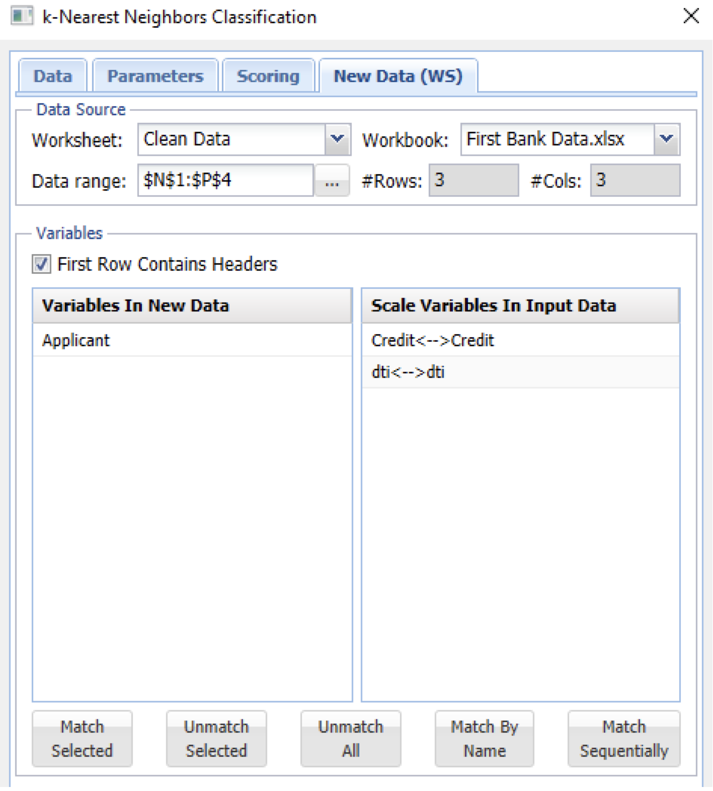
k-Nearest Neighbors Classification

| Score Training Data | Score Validation Data | Score Test Data |
|---|---|---|
| □ Detailed Report | □ Detailed Report | □ Detailed Report |
| □ Summary Report | □ Summary Report | □ Summary Report |
| □ Lift Charts | □ Lift Charts | □ Lift Charts |

Score New Data

☑ In Worksheet □ In Database

On the *New Data (WS)* tab, specify the location for the new data (N1:P4) and match the variables by name by clicking the *Match by Name* button.



Data Source

Worksheet: Clean Data

Workbook: First Bank Data.xlsx

Data range: $N$1:$P$4

#Rows: 3

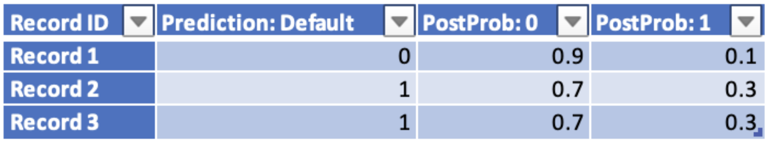
#Cols: 3

Variables

☑ First Row Contains Headers

| Variables In New Data | Scale Variables In Input Data |
|---|---|
| Applicant | Credit<-->Credit |
| | dti<-->dti |

Running the algorithm yields the results shown below. The first applicant is classified as not likely to default (probability = 0.1), while the second and third applicant are classified as likely to default (probability = 0.3).



| Record ID | Prediction: Default | PostProb: 0 | PostProb: 1 |
|---|---|---|---|
| Record 1 | 0 | 0.9 | 0.1 |
| Record 2 | 1 | 0.7 | 0.3 |
| Record 3 | 1 | 0.7 | 0.3 |

# CHAPTER 2
# OVERVIEW OF THE ANALYSIS PROCESS

## Review Questions

2.1–1 Unsecured loans don't require pledging an asset such as a house or car as collateral. Therefore, higher interest rates are typically charged for an unsecured loan as compared to a mortgage or car loan.

2.1–2 The credit score reflects the individual's payment history, debt, age, number of credit inquiries, and more.

2.2–1 The team performs a detailed technical analysis of the problem and then presents recommendations to management.

2.2–2 Management evaluates the study and its recommendations, takes into account a variety of intangible factors, and makes the final decision based on its best judgment.

2.2–3 A decision model includes (1) the individual decisions to be made, (2) the overall objective for the problem, and (3) the constraints on what can be done.

2.2–4 Common objectives include maximizing long-run profit, minimize costs, maintain stable profits, increase market share, provide for product diversification, maintain stable prices, improve worker morale, maintain family control of the business, and increase company prestige.

2.2–5 The policy has been to set the interest rate based simply on the applicant's credit score, with higher rates charged for customers with lower credit scores.

2.2–6 How often do these loans end in default? What types of customers have been more likely to default? How profitable have these loans been for First Bank?

2.2–7 It is essential to be able to predict the likelihood of default.

2.2–8 This prediction would be based upon the various characteristics of the applicant, such as their credit score, income, and debt.

2.2–9 What interest rates should be charged for different categories of customers? What is the best portfolio of loans that would provide the best tradeoff between risk and return?

2.3–1 (Step 1) gather and organize relevant data, (Step 2) clean the data, (Step 3) explore the data, (Step 4) Communicate performance information using data visualization.

2.3–2 The data surge has been a result of sophisticated computer tracking of all of the organization's internal transactions. The data also can come flooding in from sources such as web traffic, social networks, sensors of various types, and captures of audio and video recordings.

2.3–3 A numerical variable is a variable that takes on numerical values. A categorical variable can take on only a small number of values that represent the few possible categories.

2.3–4 Types of errors that are corrected during data cleaning include missing data, improperly formatted data, and duplicated data.

2.3–5 ETL is an abbreviation of *Extract. Transform,* and *Load.* The *Extract step* involves extracting the relevant data. The *Transform step* is used to “clean up” the data. The *Load step* involves moving the transformed data into a target data store.

2.3–6 The KDD process overlaps with the ETL process, but then it goes further by focusing on exploring the data to discover useful knowledge.

2.3–7 Summary statistics, such as provided by performance metrics, are useful for *quantifying* various characteristics of individual variables and the relationship between pairs of variables.

2.3–8 Sorting and filtering the data can help find outliers or mis-entered data.

2.3–9 Charts can help *visualize* the data and expose characteristics of a variable or the relationship between a pair of variables that aren’t visible through summary statistics alone.

2.3–10 The goal of data visualization is to communicate the implications of data clearly and efficiently to managers and other users through the careful selection of the most effective visual graphics.

2.3–11 The data familiarization objective is to explore the data in order to become thoroughly familiar with all of the relevant data.

2.3–12 The communication objective is to reframe raw data to make it easily understandable and meaningful to managers, investors, and other stakeholders.

2.3–13 The ultimate goal of descriptive analytics is to better understand both what has been happening in the past and what is happening now in real time, and then to develop reports that describe these understandings in the most helpful way for a wide audience.

2.4–1 (Step 1) develop a model, (Step 2) partition the data, (Step 3) test and refine the current model, (Step 4) repeat steps 1–3 for several models, and then choose the best, (Step 5) implementation.

2.4–2 Data mining is the process of finding anomalies, patterns, and correlations within large data sets to predict outcomes.

2.4–3 Some marketing applications of predictive analytics are as follows. Given the data about past sales of various products, business firms typically want the best forecasts of the future sales of these products to guide future production plans. Similarly, given information about the firm’s customers, the marketing department commonly wants to develop a marketing campaign that will particularly appeal to a certain group of these customers and lead to future sales.

2.4–4 The k-nearest-neighbor algorithm uses the observed behavior of the *k* nearest neighbors (the *k* customers whose characteristics are most similar to the prospective customer) to predict what the prospective customer will do.

2.4–5 A model that predicts a numeric outcome is referred to as a prediction model. A model that attempts to predict a yes-or-no outcome (or one of several possible outcomes) is referred to as classification model.

2.4–6 The records for the past representatives (up to the present time) are referred to as the historical records and the one for the new representative (the one for whom a prediction is needed) is called the predictor record.

2.4–7 The variable we are trying to predict is the outcome variable. The variables representing the characteristics of the records that are used to make the prediction are referred to as the predictor variables.

2.4–8 Correlation between predictor variables indicates very strong information overlap. Generally, when variables are strongly correlated, they should not *both* be included as predictor variables. A strong correlation between a particular variable and the outcome variable suggests that this particular variable may have strong predictive power for the outcome and therefore is a good candidate to be a predictor variable.

2.4–9 A model that overfits to the historical data will incorporate and account for too much of the noise, potentially to the point of missing the main signal, and therefore may make poor predictions with new data.

2.4–10 Partitioning the data into training data and validation data can be used to validate a model. Since the validation data are separate from the training data (and thus new to the model), this allows us to see how the model performs on “new” data. The validation partition can also be used as part of the process of building the model, such as choosing parameters like *k* in the k-nearest-neighbor algorithm.

2.4–11 The *specificity* is the ability to correctly predict a negative outcome, while the *sensitivity* is the ability to correctly predict a positive outcome.

2.5–1 A decision model includes decision variables, an objective function, and constraints.

2.5–2 What-if analysis (or sensitivity analysis) is used to analyze how the solution derived from the model would change (if at all) if the value assigned to a parameter were to be changed to other plausible values.

2.5–3 In a linear programming model, the mathematical functions appearing in both the objective function and the constraints are all linear functions. (A *linear function* is a mathematical expression that consists of a sum of terms where each term is simply a constant times a single variable.)

2.5–4 One advantage is that a decision model describes a problem much more concisely. This tends to make the overall structure of the problem more comprehensible, and it helps to reveal important cause-and-effect relationships. It also facilitates dealing with the problem in its entirety and considering all its interrelationships simultaneously. Finally, a decision model forms a bridge to the use of mathematical techniques and computers to analyze the problem.

2.5–5 A model is necessarily an abstract idealization of the problem, so approximations and simplifying assumptions generally are required if the model is to be *tractable* (capable of being solved). Therefore, care must be taken to ensure that the model remains a valid representation of the problem.

2.5–6 The process of model enrichment begins with a very simple version and then move in evolutionary fashion toward more elaborate models that more nearly reflect the complexity of the real problem.

2.5–7 First, they need to determine the interest rates to charge potential customers whose loan applications have been approved. Second, given the available funding, what is the best mix of unsecured and secured loans to include within First Bank's portfolio?

2.5–8 An optimal solution is the best feasible solution according to the objective function. However, since the model necessarily is an idealized rather than an exact representation of the real problem, the optimal solution may not be the best possible solution that could have been implemented for the real problem.

2.5–9 Satisficing is seeking a solution that is "good enough" for the problem at hand, rather than the solution that optimizes a measure of performance. Goals may be set to establish minimum satisfactory levels of performance or include constraints that require the goals to achieve at least their minimum satisfactory levels of performance.

2.5–10 Heuristic procedures are intuitively designed procedures that do not guarantee an optimal solution but find a good suboptimal solution. Metaheuristics provide both a general structure and strategy guidelines for designing a specific heuristic procedure to fit a particular kind of problem.

2.5–11 An optimal solution for the original model may be far from ideal for the real problem.

2.5–12 A parameters is said to be a sensitive parameter if (without changing any other parameter values) its value cannot be changed without changing the optimal solution. Identifying the sensitive parameters is important, because this identifies the parameters whose value must be assigned with special care to avoid distorting the output of the model.

2.5–13 The first version of a large decision model inevitably contains many flaws. Some relevant factors or interrelationships may not have been incorporated into the model, and some parameters frequently have not been estimated correctly. Therefore, before you use the model, it must be thoroughly tested to try to identify and correct as many flaws as possible.

2.5–14 A retrospective test involves using historical data to reconstruct the past and then determining how well the model and the resulting solution would have performed if they had been used. Considerable evidence can be gathered regarding how well the model predicts the relative effects of alternative courses of actions. However, it uses the same data that guided the formulation of the model. The crucial question is whether the past is truly representative of the future. If it is not, then the model might perform quite differently in the future than it would have in the past.

2.5–15 Documenting the process helps to increase confidence in the model for subsequent users. Furthermore, if concerns arise in the future about the model, this documentation will be helpful in diagnosing where problems may lie.

2.5–16 Databases and management information systems may provide up-to-date input for the model each time it is used. An *interactive* computer-based system called a decision support system is installed to help managers use data and models to support (rather than replace) their decision making as needed. Additional software may generate *managerial reports* (in the language of management) that interpret the output of the model and its implications for application.

2.5–17 It is important for the study team to participate in launching this phase, both to make sure that model solutions are accurately translated to an operating procedure and to rectify any flaws that are uncovered in the solutions.

2.5–18 First, the study team gives operating management a careful explanation of the new system to be adopted and how it relates to operating realities. Next, these two parties share the responsibility for developing the procedures required to put this system into operation. Operating management then makes sure that a detailed indoctrination is given to the personnel involved.

## Problems

2.1 The core component of the ad-serving algorithm is a classification model that computes, for each candidate ad video, the probability that it will generate a profitable action (i.e., a view, click, or installation). A two-stage approach incorporates budget restrictions and user-fatigue issues. See pages 457-460 of the article for details.

2.2 The core component of the ad-serving algorithm is a classification model that computes, for each candidate ad video, the probability that it will generate a profitable action (i.e., a view, click, or installation). A two-stage approach incorporates budget restrictions and user-fatigue issues. See pages 457-460 of the article for details.

The financial benefits that resulted from this study include savings of $40 million in 2001 and of $5 million in 2002. The savings for any major disruption have been between $1 and $5 million. The new system enabled Continental Airlines to operate in an efficient and cost-effective manner in case of disruptions. The time to recover and the costs associated with disruptions are reduced. What-if analysis allowed the company to evaluate various scenarios in short periods of time. Since the complete reliable data can be generated quickly, the company reacts to facts rather than forecasts. These improvements in handling irregularities resulted in better and more reliable service and hence happier customers.

2.3 *a.* False. The study team works in an advisory capacity, so the goal is to make recommendations to management for how to do this.

*b.* False. Management tends to adopt the goal of satisfactory profits combined with other objectives.

*c.* True.

2.4 *a.* False. This is true for business analysts who specialize mainly in business analytics instead.

*b.* False. Data mining is the process of finding anomalies, patterns, and correlations within large data sets to predict outcomes.

*c.* True.

2.5 *a.* False. The goal is to understand what has been happening to date and then to develop insightful reports that describe these understandings.

*b.* False. The goal is to communicate the implications of data clearly and efficiently through using effective visual graphics.

*c.* False. The focus is on using data to predict future events, trends, or behaviors.

2.6 *a.* True.

*b.* True.

*c.* True.

2.7 *a.* False. A good approach is to begin with a very simple version and then move in evolutionary fashion toward more elaborate models that more nearly reflect the complexity of the real problem.

*b.* True.

*c.* False. It often is not possible to obtain an optimal solution. The test of the practical success of a study should be whether it provides a better guide for action than can be obtained by other means.

2.8 *a.* False. It addresses some questions about what would happen to the optimal solution if different assumptions were made about future conditions.

*b.* True.

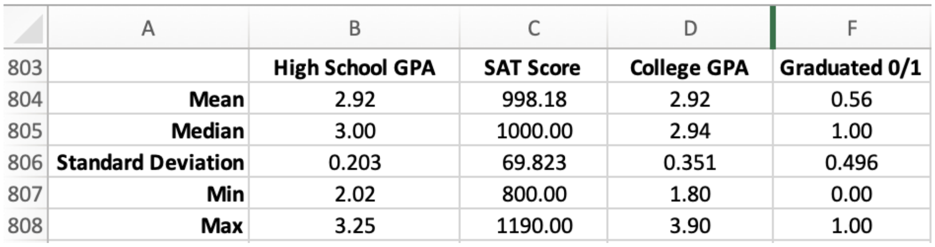
*c.* False. It is important for the study team to participate in launching the implementation as well.

2.9 *a.* Student 250 missing Graduated data.
Student 506 missing College GPA data.
Student 616 missing High School GPA data.
Student 655 missing SAT data.

*b.* Student 190 SAT entered as 8.90. Possibly a stray decimal point (890).
Student 333 Graduated entered as Y. Should probably be Yes.
Student 421 High School GPA entered as 287. Possibly should be 2.87.
Student 547 Graduated entered as N. Should probably be No.

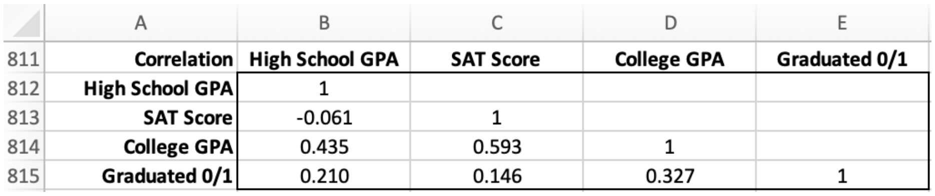
*c.* Enter =IF(E2="Yes",1,0) and fill down.

2.10 *a.* Summary Statistics:



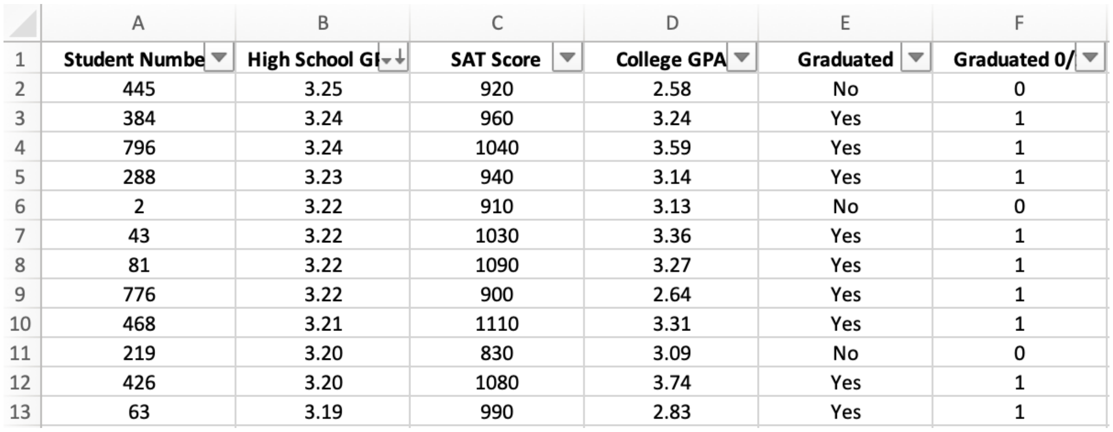
| | A | B | C | D | F |
|---|---|---|---|---|---|
| 803 | | **High School GPA** | **SAT Score** | **College GPA** | **Graduated 0/1** |
| 804 | **Mean** | 2.92 | 998.18 | 2.92 | 0.56 |
| 805 | **Median** | 3.00 | 1000.00 | 2.94 | 1.00 |
| 806 | **Standard Deviation** | 0.203 | 69.823 | 0.351 | 0.496 |
| 807 | **Min** | 2.02 | 800.00 | 1.80 | 0.00 |
| 808 | **Max** | 3.25 | 1190.00 | 3.90 | 1.00 |

*b.* Correlation Table:



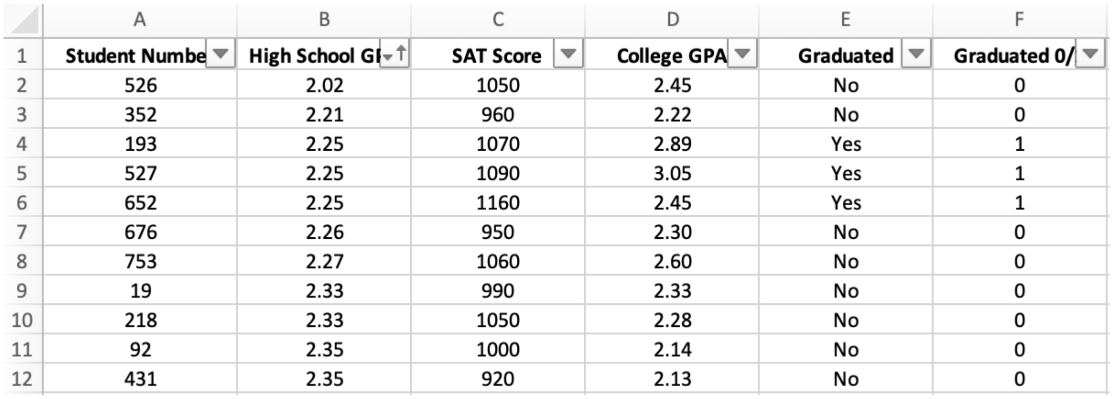
| | A | B | C | D | E |
|---|---|---|---|---|---|
| 811 | Correlation | High School GPA | SAT Score | College GPA | Graduated 0/1 |
| 812 | High School GPA | 1 | | | |
| 813 | SAT Score | -0.061 | 1 | | |
| 814 | College GPA | 0.435 | 0.593 | 1 | |
| 815 | Graduated 0/1 | 0.210 | 0.146 | 0.327 | 1 |

2.11 *a.* Most with the highest *High School GPA* have graduated



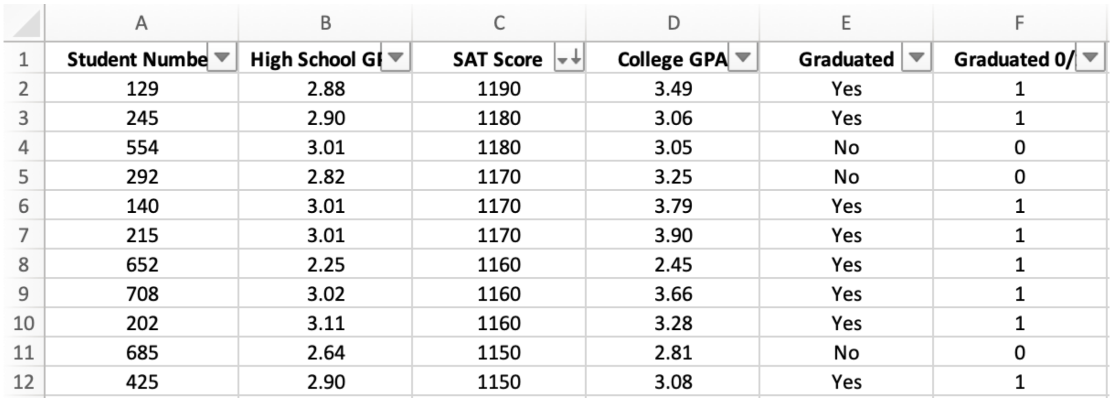
| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | Student Numbe | High School GI | SAT Score | College GPA | Graduated | Graduated 0/ |
| 2 | 445 | 3.25 | 920 | 2.58 | No | 0 |
| 3 | 384 | 3.24 | 960 | 3.24 | Yes | 1 |
| 4 | 796 | 3.24 | 1040 | 3.59 | Yes | 1 |
| 5 | 288 | 3.23 | 940 | 3.14 | Yes | 1 |
| 6 | 2 | 3.22 | 910 | 3.13 | No | 0 |
| 7 | 43 | 3.22 | 1030 | 3.36 | Yes | 1 |
| 8 | 81 | 3.22 | 1090 | 3.27 | Yes | 1 |
| 9 | 776 | 3.22 | 900 | 2.64 | Yes | 1 |
| 10 | 468 | 3.21 | 1110 | 3.31 | Yes | 1 |
| 11 | 219 | 3.20 | 830 | 3.09 | No | 0 |
| 12 | 426 | 3.20 | 1080 | 3.74 | Yes | 1 |
| 13 | 63 | 3.19 | 990 | 2.83 | Yes | 1 |

Many with the lowest *High School GPA* did not graduate



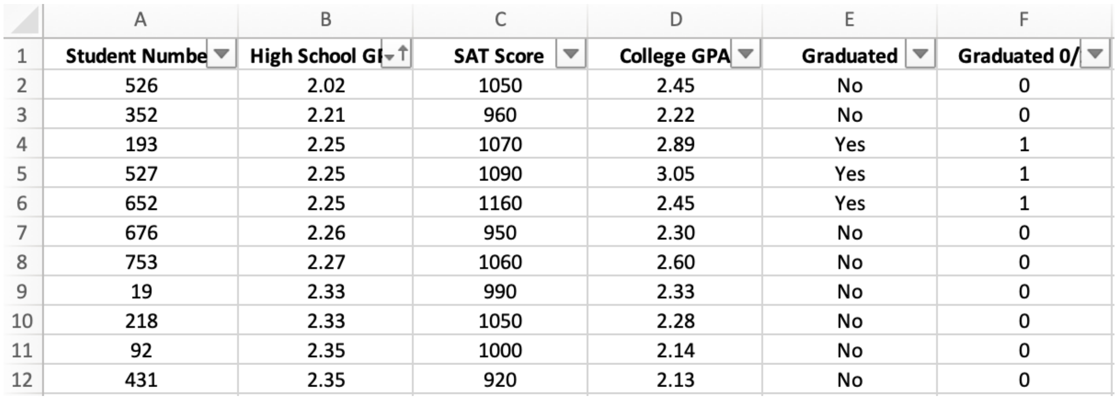
| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | Student Numbe | High School GI | SAT Score | College GPA | Graduated | Graduated 0/ |
| 2 | 526 | 2.02 | 1050 | 2.45 | No | 0 |
| 3 | 352 | 2.21 | 960 | 2.22 | No | 0 |
| 4 | 193 | 2.25 | 1070 | 2.89 | Yes | 1 |
| 5 | 527 | 2.25 | 1090 | 3.05 | Yes | 1 |
| 6 | 652 | 2.25 | 1160 | 2.45 | Yes | 1 |
| 7 | 676 | 2.26 | 950 | 2.30 | No | 0 |
| 8 | 753 | 2.27 | 1060 | 2.60 | No | 0 |
| 9 | 19 | 2.33 | 990 | 2.33 | No | 0 |
| 10 | 218 | 2.33 | 1050 | 2.28 | No | 0 |
| 11 | 92 | 2.35 | 1000 | 2.14 | No | 0 |
| 12 | 431 | 2.35 | 920 | 2.13 | No | 0 |

*b.* Most with the highest *SAT Scores* have graduated



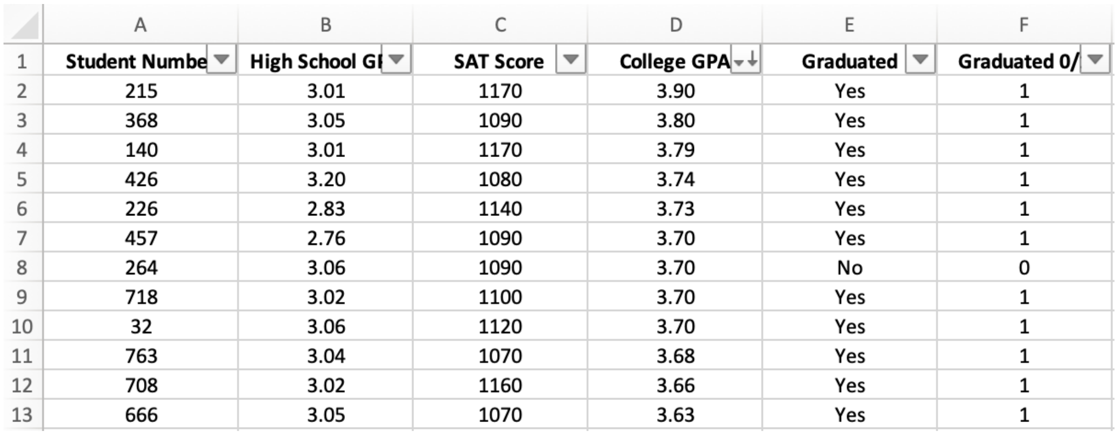
| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | Student Numbe | High School GI | SAT Score | College GPA | Graduated | Graduated 0/ |
| 2 | 129 | 2.88 | 1190 | 3.49 | Yes | 1 |
| 3 | 245 | 2.90 | 1180 | 3.06 | Yes | 1 |
| 4 | 554 | 3.01 | 1180 | 3.05 | No | 0 |
| 5 | 292 | 2.82 | 1170 | 3.25 | No | 0 |
| 6 | 140 | 3.01 | 1170 | 3.79 | Yes | 1 |
| 7 | 215 | 3.01 | 1170 | 3.90 | Yes | 1 |
| 8 | 652 | 2.25 | 1160 | 2.45 | Yes | 1 |
| 9 | 708 | 3.02 | 1160 | 3.66 | Yes | 1 |
| 10 | 202 | 3.11 | 1160 | 3.28 | Yes | 1 |
| 11 | 685 | 2.64 | 1150 | 2.81 | No | 0 |
| 12 | 425 | 2.90 | 1150 | 3.08 | Yes | 1 |

Many with the lowest *SAT Scores* did not graduate



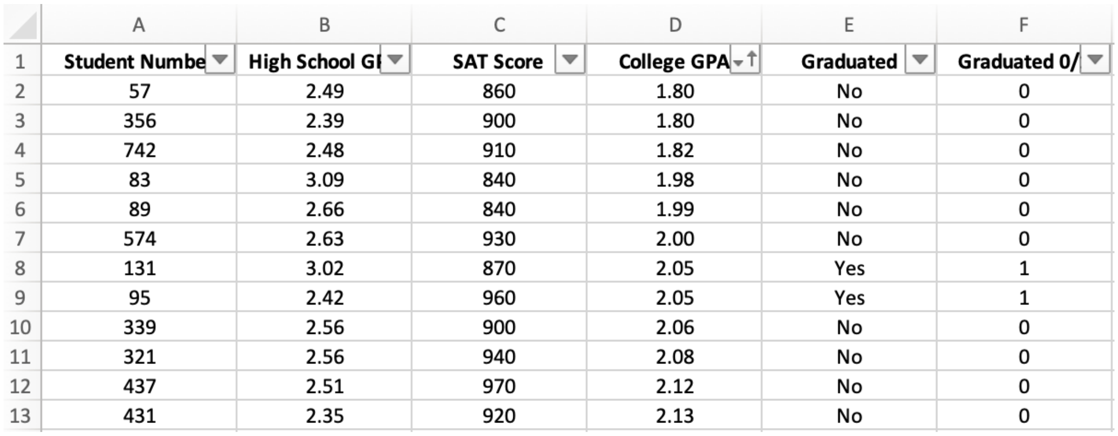
| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | Student Numbe | High School GI | SAT Score | College GPA | Graduated | Graduated 0/ |
| 2 | 526 | 2.02 | 1050 | 2.45 | No | 0 |
| 3 | 352 | 2.21 | 960 | 2.22 | No | 0 |
| 4 | 193 | 2.25 | 1070 | 2.89 | Yes | 1 |
| 5 | 527 | 2.25 | 1090 | 3.05 | Yes | 1 |
| 6 | 652 | 2.25 | 1160 | 2.45 | Yes | 1 |
| 7 | 676 | 2.26 | 950 | 2.30 | No | 0 |
| 8 | 753 | 2.27 | 1060 | 2.60 | No | 0 |
| 9 | 19 | 2.33 | 990 | 2.33 | No | 0 |
| 10 | 218 | 2.33 | 1050 | 2.28 | No | 0 |
| 11 | 92 | 2.35 | 1000 | 2.14 | No | 0 |
| 12 | 431 | 2.35 | 920 | 2.13 | No | 0 |

*c.* Essentially all with the highest *College GPA* have graduated



| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | Student Numbe | High School GI | SAT Score | College GPA | Graduated | Graduated 0/ |
| 2 | 215 | 3.01 | 1170 | 3.90 | Yes | 1 |
| 3 | 368 | 3.05 | 1090 | 3.80 | Yes | 1 |
| 4 | 140 | 3.01 | 1170 | 3.79 | Yes | 1 |
| 5 | 426 | 3.20 | 1080 | 3.74 | Yes | 1 |
| 6 | 226 | 2.83 | 1140 | 3.73 | Yes | 1 |
| 7 | 457 | 2.76 | 1090 | 3.70 | Yes | 1 |
| 8 | 264 | 3.06 | 1090 | 3.70 | No | 0 |
| 9 | 718 | 3.02 | 1100 | 3.70 | Yes | 1 |
| 10 | 32 | 3.06 | 1120 | 3.70 | Yes | 1 |
| 11 | 763 | 3.04 | 1070 | 3.68 | Yes | 1 |
| 12 | 708 | 3.02 | 1160 | 3.66 | Yes | 1 |
| 13 | 666 | 3.05 | 1070 | 3.63 | Yes | 1 |

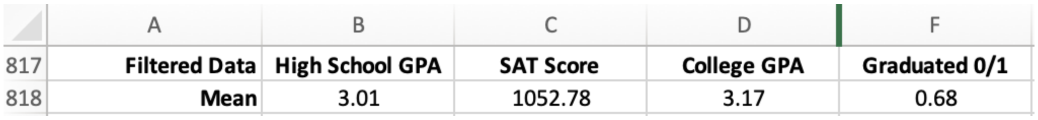
Most with the lowest *College GPA* did not graduate



| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | Student Numbe | High School GI | SAT Score | College GPA | Graduated | Graduated 0/ |
| 2 | 57 | 2.49 | 860 | 1.80 | No | 0 |
| 3 | 356 | 2.39 | 900 | 1.80 | No | 0 |
| 4 | 742 | 2.48 | 910 | 1.82 | No | 0 |
| 5 | 83 | 3.09 | 840 | 1.98 | No | 0 |
| 6 | 89 | 2.66 | 840 | 1.99 | No | 0 |
| 7 | 574 | 2.63 | 930 | 2.00 | No | 0 |
| 8 | 131 | 3.02 | 870 | 2.05 | Yes | 1 |
| 9 | 95 | 2.42 | 960 | 2.05 | Yes | 1 |
| 10 | 339 | 2.56 | 900 | 2.06 | No | 0 |
| 11 | 321 | 2.56 | 940 | 2.08 | No | 0 |
| 12 | 437 | 2.51 | 970 | 2.12 | No | 0 |
| 13 | 431 | 2.35 | 920 | 2.13 | No | 0 |

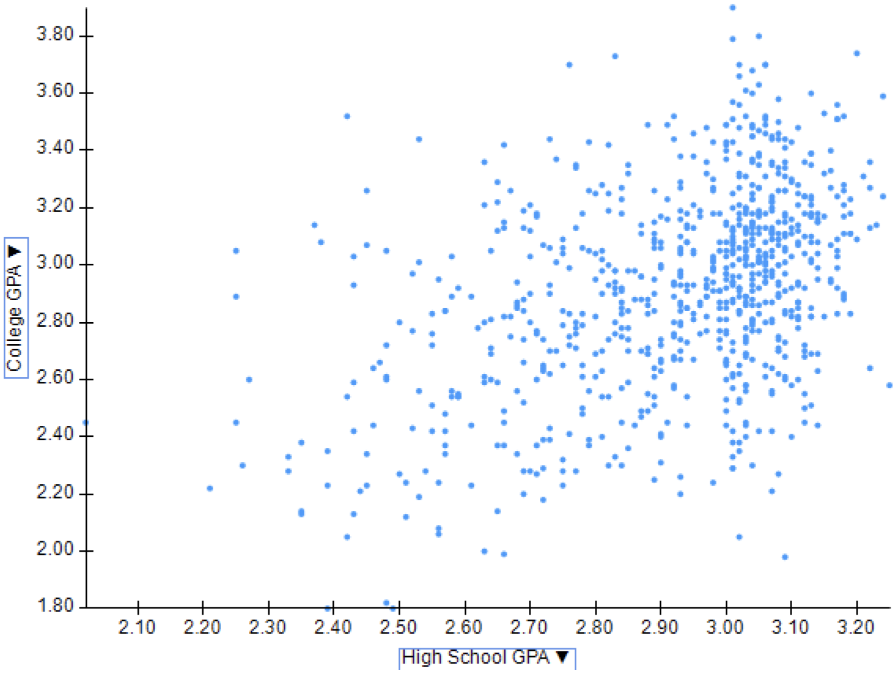
2.12 *a.* Enter =AGGREGATE(1,5,B2:B801) and fill right.

*b.* With *High School GPA* filtered to only show ≥2.80, and *SAT Score* filtered to only show ≥1000, the mean graduation rate is 0.68.

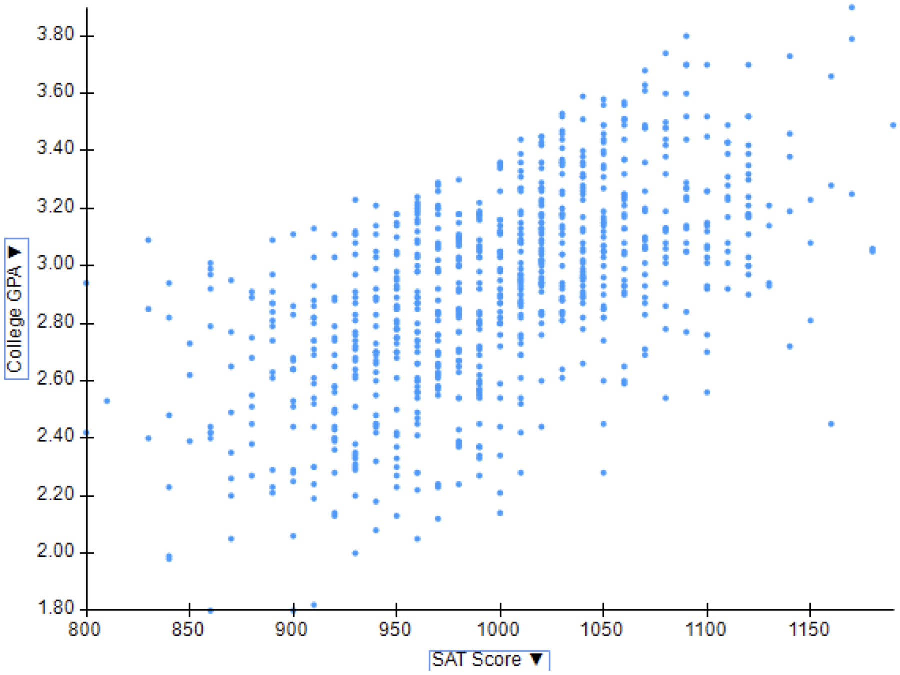


| | A | B | C | D | F |
|---|---|---|---|---|---|
| 817 | Filtered Data | High School GPA | SAT Score | College GPA | Graduated 0/1 |
| 818 | Mean | 3.01 | 1052.78 | 3.17 | 0.68 |

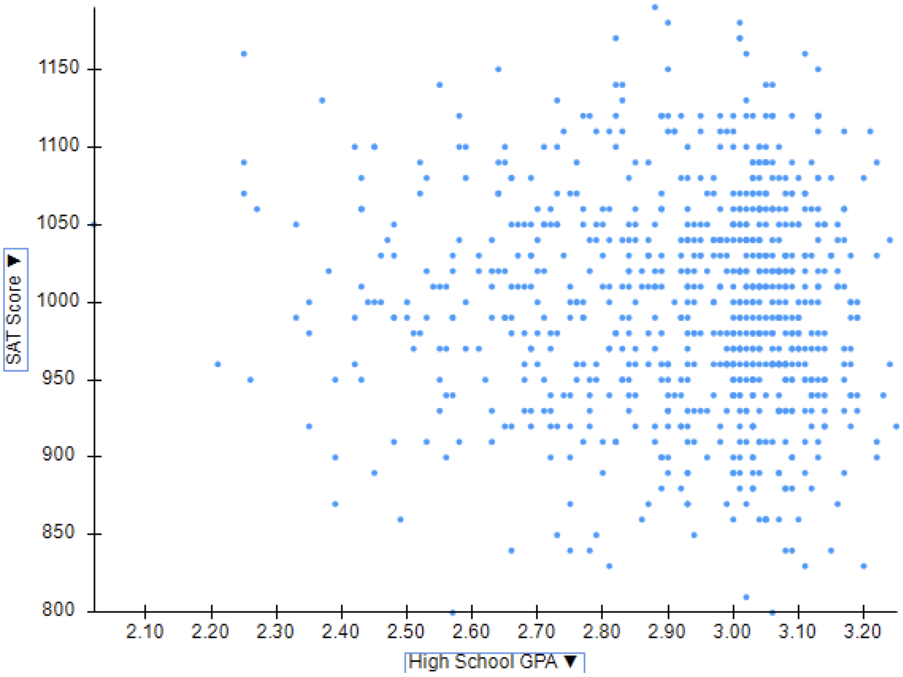
2.13 *a.* A positive correlation between *College GPA* and *High School GPA* is observed.



*b.* A positive correlation between *College GPA* and *SAT Score* is observed.



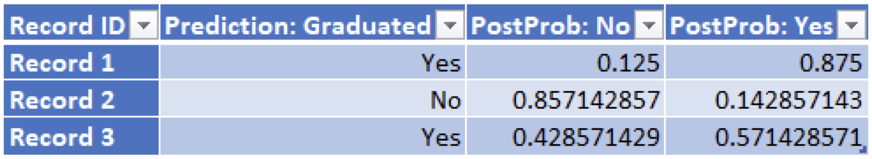
*c.* Very little correlation between *SAT Score* and *High School GPA* is observed.



2.14 *a.* Classification: Yes (likely to graduate). Probability of graduation ≈ 0.875.

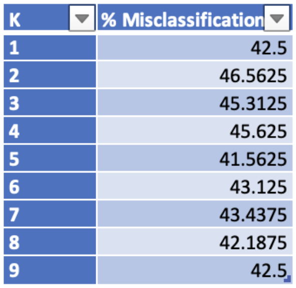
*b.* Classification: No (not likely to graduate). Probability of graduation ≈ 0.143.

*c.* Classification: Yes (likely to graduate). Probability of graduation ≈ 0.571.



| Record ID | Prediction: Graduated | PostProb: No | PostProb: Yes |
|---|---|---|---|
| Record 1 | Yes | 0.125 | 0.875 |
| Record 2 | No | 0.857142857 | 0.142857143 |
| Record 3 | Yes | 0.428571429 | 0.571428571 |

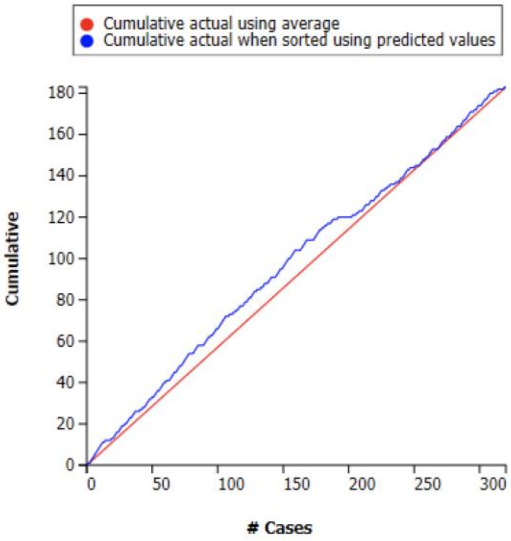
2.15 *a.* Using *k* = 5 gives the highest accuracy.



| K | % Misclassification |
|---|---|
| 1 | 42.5 |
| 2 | 46.5625 |
| 3 | 45.3125 |
| 4 | 45.625 |
| 5 | 41.5625 |
| 6 | 43.125 |
| 7 | 43.4375 |
| 8 | 42.1875 |
| 9 | 42.5 |

*b.* Accuracy 58.4%, Specificity 49.6%, Sensitivity 65.0%

*c.* Lift Chart

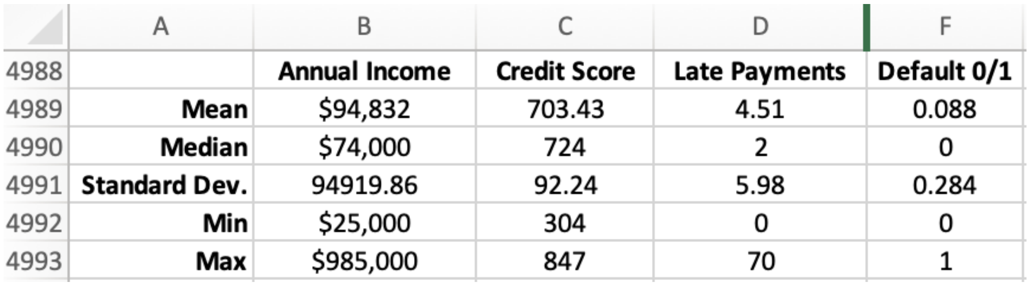


2.16 *a.* Loan 689 is missing *Annual Income* data.
Loan 1157 is missing *Default* data.
Loan 2324 is missing *Late Payments* data.

*b.* Loan 1455 has *Annual Income* entered as \$83. Perhaps should be \$83,000.
Loan 2457 has *Credit Score* entered as 5.40. Perhaps a stray decimal (540).
Loan 4504 has *Default* entered as N. Perhaps should be No.

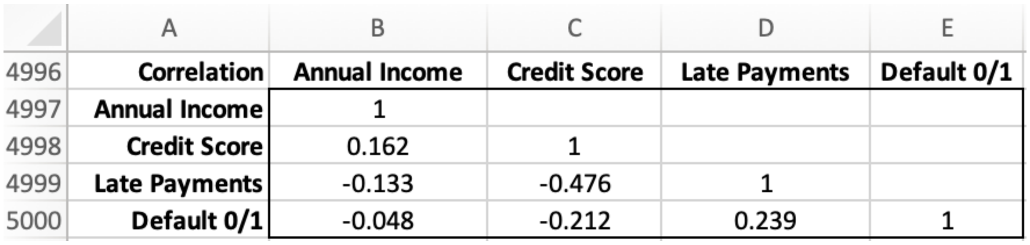
*c.* Enter =IF(E2="Yes",1,0) and fill down.

2.17 *a.* Summary Statistics



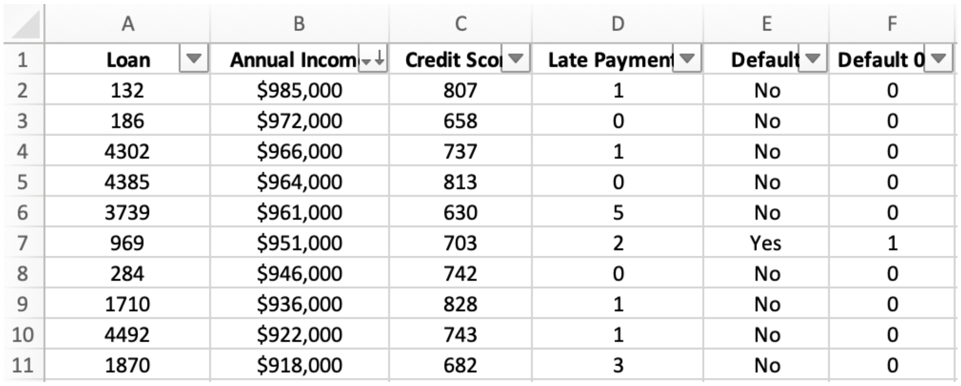
| | A | B | C | D | F |
|---|---|---|---|---|---|
| 4988 | | Annual Income | Credit Score | Late Payments | Default 0/1 |
| 4989 | Mean | $94,832 | 703.43 | 4.51 | 0.088 |
| 4990 | Median | $74,000 | 724 | 2 | 0 |
| 4991 | Standard Dev. | 94919.86 | 92.24 | 5.98 | 0.284 |
| 4992 | Min | $25,000 | 304 | 0 | 0 |
| 4993 | Max | $985,000 | 847 | 70 | 1 |

*b.* Correlation Table



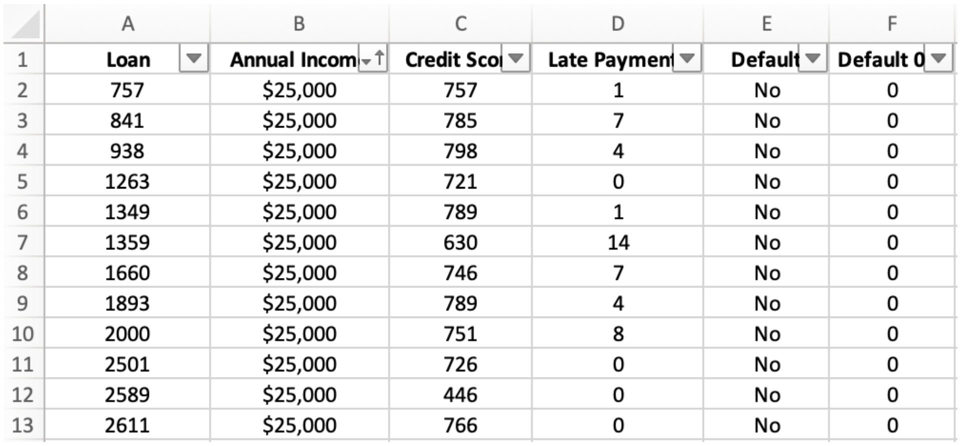
| | A | B | C | D | E |
|---|---|---|---|---|---|
| 4996 | Correlation | Annual Income | Credit Score | Late Payments | Default 0/1 |
| 4997 | Annual Income | 1 | | | |
| 4998 | Credit Score | 0.162 | 1 | | |
| 4999 | Late Payments | -0.133 | -0.476 | 1 | |
| 5000 | Default 0/1 | -0.048 | -0.212 | 0.239 | 1 |

2.18 *a.* Most with the highest *Annual Income* did not default.



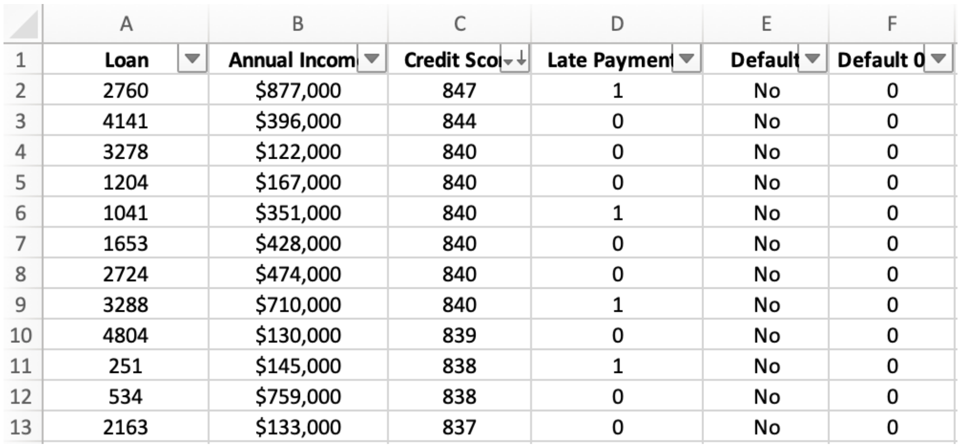
| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | Loan | Annual Incom | Credit Sco | Late Paymen | Default | Default 0 |
| 2 | 132 | $985,000 | 807 | 1 | No | 0 |
| 3 | 186 | $972,000 | 658 | 0 | No | 0 |
| 4 | 4302 | $966,000 | 737 | 1 | No | 0 |
| 5 | 4385 | $964,000 | 813 | 0 | No | 0 |
| 6 | 3739 | $961,000 | 630 | 5 | No | 0 |
| 7 | 969 | $951,000 | 703 | 2 | Yes | 1 |
| 8 | 284 | $946,000 | 742 | 0 | No | 0 |
| 9 | 1710 | $936,000 | 828 | 1 | No | 0 |
| 10 | 4492 | $922,000 | 743 | 1 | No | 0 |
| 11 | 1870 | $918,000 | 682 | 3 | No | 0 |

Those with the very lowest *Annual Income* did not default.



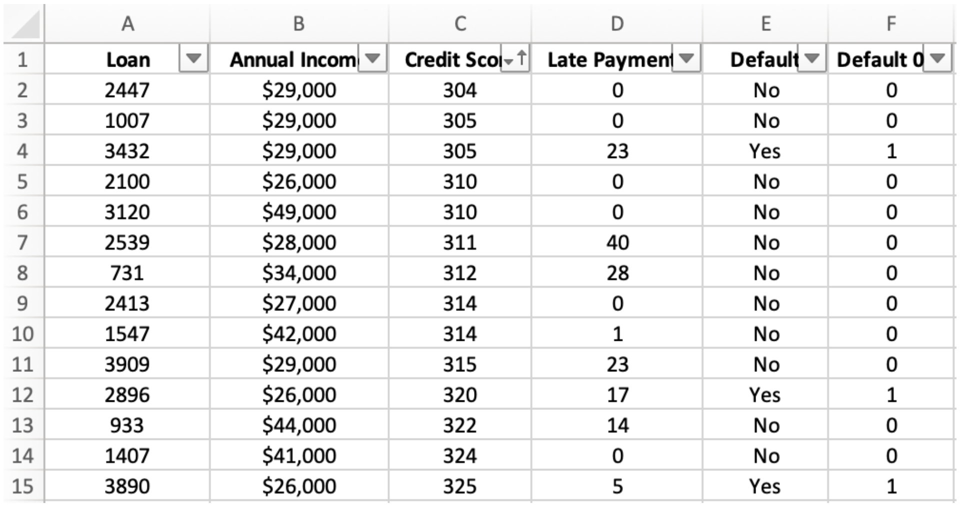
| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | Loan | Annual Incom | Credit Sco | Late Paymen | Default | Default 0 |
| 2 | 757 | $25,000 | 757 | 1 | No | 0 |
| 3 | 841 | $25,000 | 785 | 7 | No | 0 |
| 4 | 938 | $25,000 | 798 | 4 | No | 0 |
| 5 | 1263 | $25,000 | 721 | 0 | No | 0 |
| 6 | 1349 | $25,000 | 789 | 1 | No | 0 |
| 7 | 1359 | $25,000 | 630 | 14 | No | 0 |
| 8 | 1660 | $25,000 | 746 | 7 | No | 0 |
| 9 | 1893 | $25,000 | 789 | 4 | No | 0 |
| 10 | 2000 | $25,000 | 751 | 8 | No | 0 |
| 11 | 2501 | $25,000 | 726 | 0 | No | 0 |
| 12 | 2589 | $25,000 | 446 | 0 | No | 0 |
| 13 | 2611 | $25,000 | 766 | 0 | No | 0 |

*b.* Those with the very highest *Credit Scores* did not default.



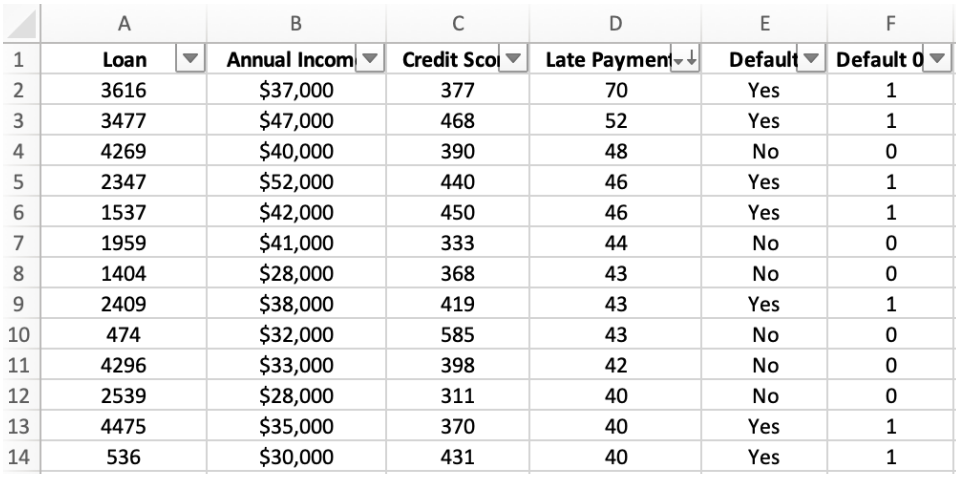
| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | Loan | Annual Incom | Credit Sco | Late Paymen | Default | Default 0 |
| 2 | 2760 | $877,000 | 847 | 1 | No | 0 |
| 3 | 4141 | $396,000 | 844 | 0 | No | 0 |
| 4 | 3278 | $122,000 | 840 | 0 | No | 0 |
| 5 | 1204 | $167,000 | 840 | 0 | No | 0 |
| 6 | 1041 | $351,000 | 840 | 1 | No | 0 |
| 7 | 1653 | $428,000 | 840 | 0 | No | 0 |
| 8 | 2724 | $474,000 | 840 | 0 | No | 0 |
| 9 | 3288 | $710,000 | 840 | 1 | No | 0 |
| 10 | 4804 | $130,000 | 839 | 0 | No | 0 |
| 11 | 251 | $145,000 | 838 | 1 | No | 0 |
| 12 | 534 | $759,000 | 838 | 0 | No | 0 |
| 13 | 2163 | $133,000 | 837 | 0 | No | 0 |

Some with the lowest *Credit Scores* defaulted, and many had a lot of late payments.



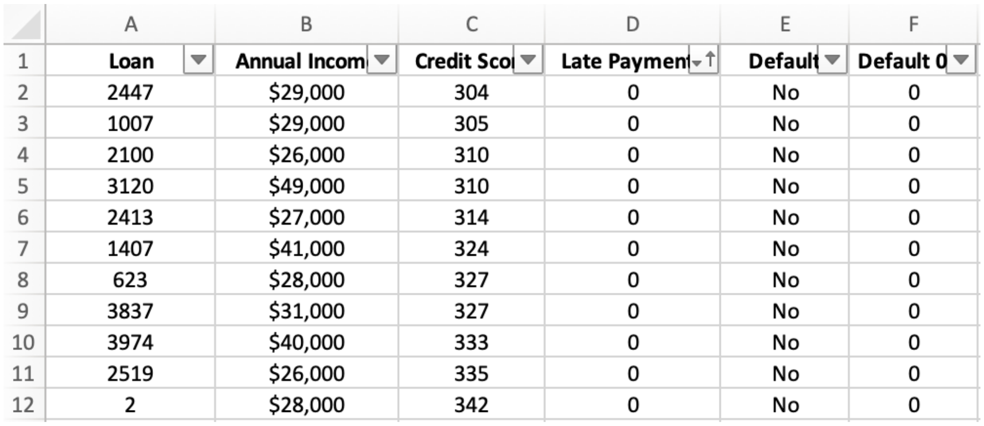
| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | Loan | Annual Incom | Credit Sco | Late Paymen | Default | Default 0 |
| 2 | 2447 | $29,000 | 304 | 0 | No | 0 |
| 3 | 1007 | $29,000 | 305 | 0 | No | 0 |
| 4 | 3432 | $29,000 | 305 | 23 | Yes | 1 |
| 5 | 2100 | $26,000 | 310 | 0 | No | 0 |
| 6 | 3120 | $49,000 | 310 | 0 | No | 0 |
| 7 | 2539 | $28,000 | 311 | 40 | No | 0 |
| 8 | 731 | $34,000 | 312 | 28 | No | 0 |
| 9 | 2413 | $27,000 | 314 | 0 | No | 0 |
| 10 | 1547 | $42,000 | 314 | 1 | No | 0 |
| 11 | 3909 | $29,000 | 315 | 23 | No | 0 |
| 12 | 2896 | $26,000 | 320 | 17 | Yes | 1 |
| 13 | 933 | $44,000 | 322 | 14 | No | 0 |
| 14 | 1407 | $41,000 | 324 | 0 | No | 0 |
| 15 | 3890 | $26,000 | 325 | 5 | Yes | 1 |

*c.* Many with the highest number of *Late Payments* ended up defaulting.



| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | Loan | Annual Incom | Credit Sco | Late Paymen | Default | Default 0 |
| 2 | 3616 | $37,000 | 377 | 70 | Yes | 1 |
| 3 | 3477 | $47,000 | 468 | 52 | Yes | 1 |
| 4 | 4269 | $40,000 | 390 | 48 | No | 0 |
| 5 | 2347 | $52,000 | 440 | 46 | Yes | 1 |
| 6 | 1537 | $42,000 | 450 | 46 | Yes | 1 |
| 7 | 1959 | $41,000 | 333 | 44 | No | 0 |
| 8 | 1404 | $28,000 | 368 | 43 | No | 0 |
| 9 | 2409 | $38,000 | 419 | 43 | Yes | 1 |
| 10 | 474 | $32,000 | 585 | 43 | No | 0 |
| 11 | 4296 | $33,000 | 398 | 42 | No | 0 |
| 12 | 2539 | $28,000 | 311 | 40 | No | 0 |
| 13 | 4475 | $35,000 | 370 | 40 | Yes | 1 |
| 14 | 536 | $30,000 | 431 | 40 | Yes | 1 |

Those with the fewest *Late Payments* generally did not default.



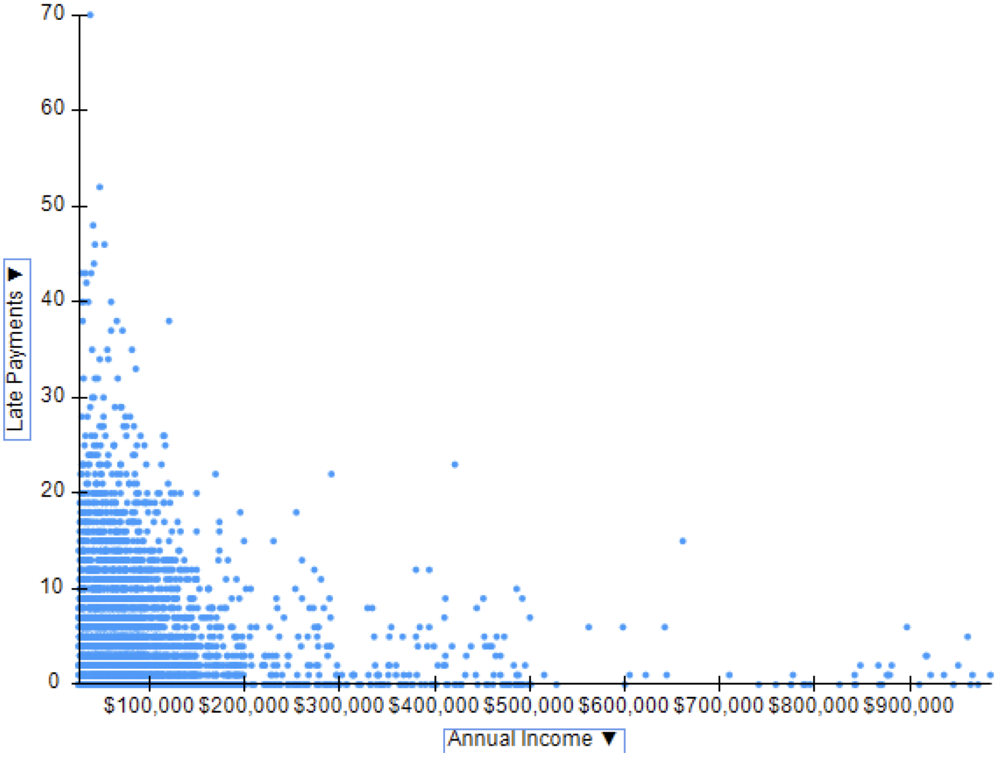
| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | Loan | Annual Incom | Credit Sco | Late Paymen | Default | Default 0 |
| 2 | 2447 | $29,000 | 304 | 0 | No | 0 |
| 3 | 1007 | $29,000 | 305 | 0 | No | 0 |
| 4 | 2100 | $26,000 | 310 | 0 | No | 0 |
| 5 | 3120 | $49,000 | 310 | 0 | No | 0 |
| 6 | 2413 | $27,000 | 314 | 0 | No | 0 |
| 7 | 1407 | $41,000 | 324 | 0 | No | 0 |
| 8 | 623 | $28,000 | 327 | 0 | No | 0 |
| 9 | 3837 | $31,000 | 327 | 0 | No | 0 |
| 10 | 3974 | $40,000 | 333 | 0 | No | 0 |
| 11 | 2519 | $26,000 | 335 | 0 | No | 0 |
| 12 | 2 | $28,000 | 342 | 0 | No | 0 |

2.19 *a.* Enter =AGGREGATE(1,5,B2:B4986) and fill right.

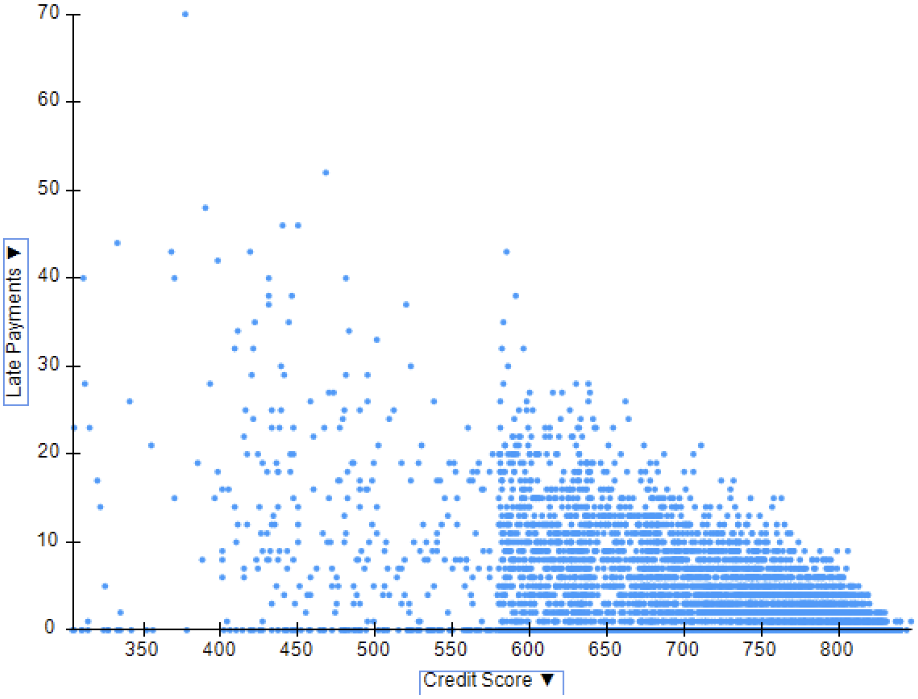
*b.* With *Annual Income* filtered to only show ≥35,000, and *Credit Score* filtered to only show ≥600, the mean default rate is 7 percent.

| Filtered Data | Annual Income | Credit Score | Late Payments | | Default 0/1 |
|---|---|---|---|---|---|
| Mean | $105,943.82 | 728.10 | 3.71 | | 0.070 |

2.20 *a.* A negative correlation between *Late Payments* and *Income* is observed.



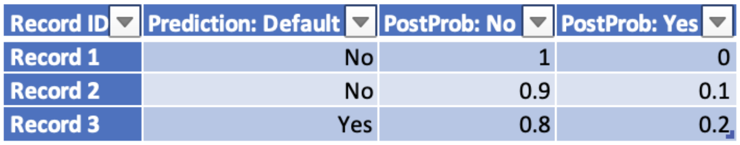
*b.* A negative correlation between *Late Payments* and *Credit Score* is observed.



2.21 *a.* Classification: No (not likely to default). Probability of default ≈ 0.

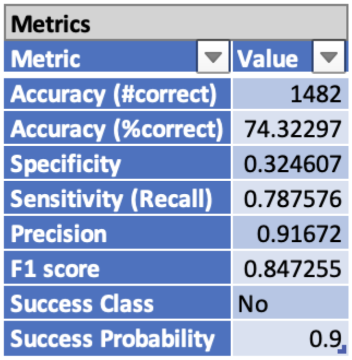
*b.* Classification: No (not likely to default). Probability of default ≈ 0.1.

*c.* Classification: Yes (likely to default). Probability of default ≈ 0.2.



| Record ID | Prediction: Default | PostProb: No | PostProb: Yes |
|---|---|---|---|
| Record 1 | No | 1 | 0 |
| Record 2 | No | 0.9 | 0.1 |
| Record 3 | Yes | 0.8 | 0.2 |

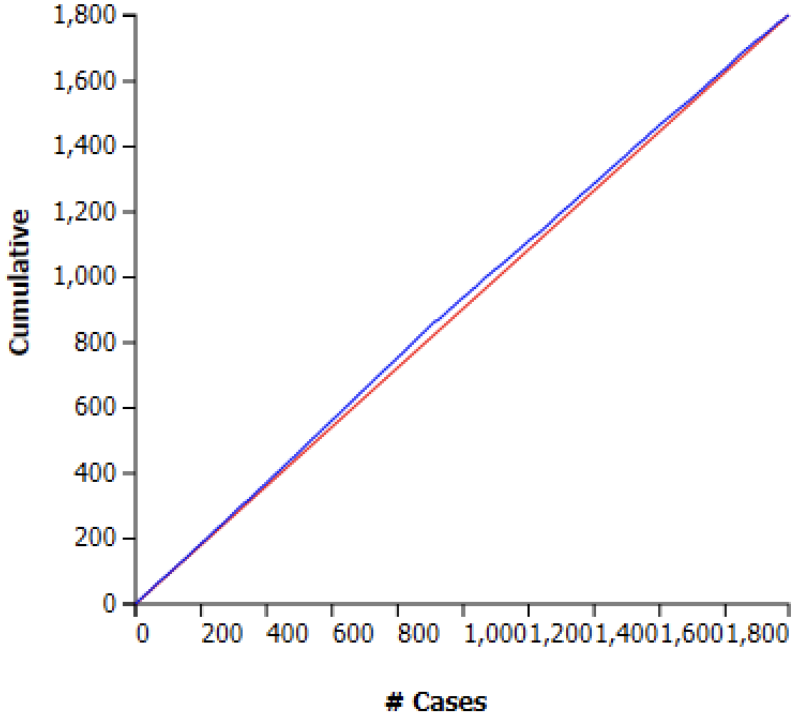
2.22 *a.* Accuracy 74.3%, Specificity 32.5%, Sensitivity 78.8%



Metrics

| Metric | Value |
|---|---|
| Accuracy (#correct) | 1482 |
| Accuracy (%correct) | 74.32297 |
| Specificity | 0.324607 |
| Sensitivity (Recall) | 0.787576 |
| Precision | 0.91672 |
| F1 score | 0.847255 |
| Success Class | No |
| Success Probability | 0.9 |

*b.* Lift Chart



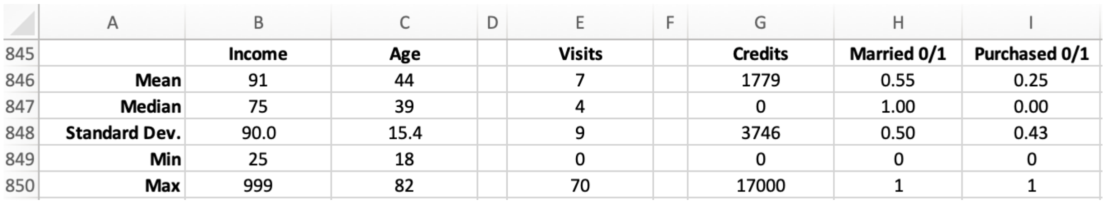
## Cases

2–1 *a.* Record 102 is missing Purchase and Credits data.
Record 287 is missing Income data.

*b.* Record 107 Income is 58234 ($58,234,000). Perhaps entered as $ instead of $thousands.
Record 299 Marital Status entered as M. Probably should be Married.
Record 354 Marital Status entered as S. Probably should be Single.
Record 457 Credits is 3. Perhaps should be 3000.

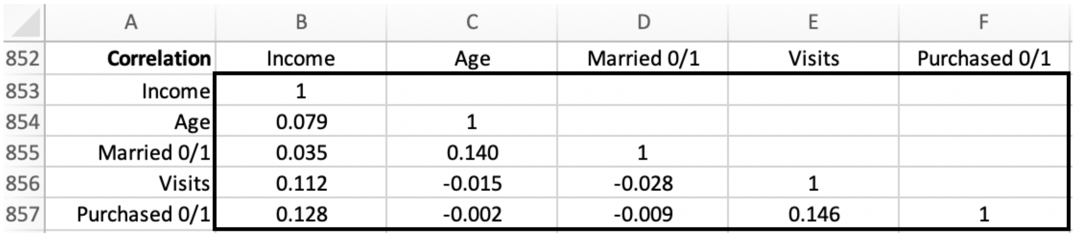
*c.* Enter =IF(D2="Married", 1, 0) and fill down.
Enter =IF(F2="Yes", 1, 0) and fill down.

*d.* Statistics for variables



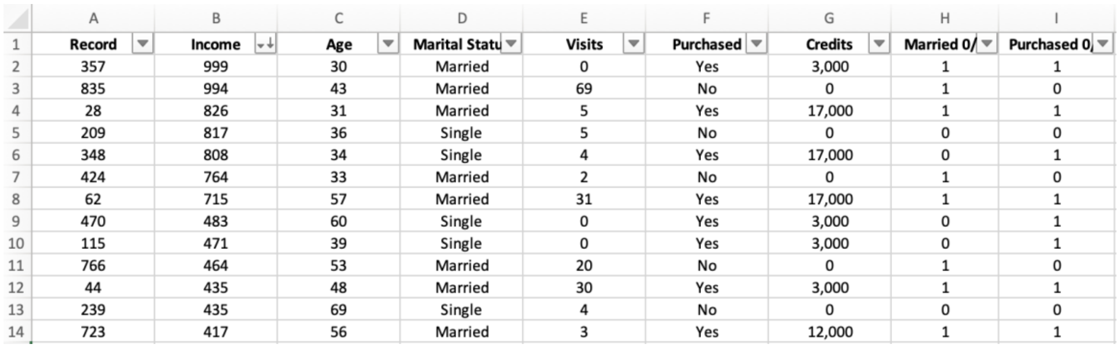
|  | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 845 |  | Income | Age |  | Visits |  | Credits | Married 0/1 | Purchased 0/1 |
| 846 | Mean | 91 | 44 |  | 7 |  | 1779 | 0.55 | 0.25 |
| 847 | Median | 75 | 39 |  | 4 |  | 0 | 1.00 | 0.00 |
| 848 | Standard Dev. | 90.0 | 15.4 |  | 9 |  | 3746 | 0.50 | 0.43 |
| 849 | Min | 25 | 18 |  | 0 |  | 0 | 0 | 0 |
| 850 | Max | 999 | 82 |  | 70 |  | 17000 | 1 | 1 |

*e.* Correlation Table



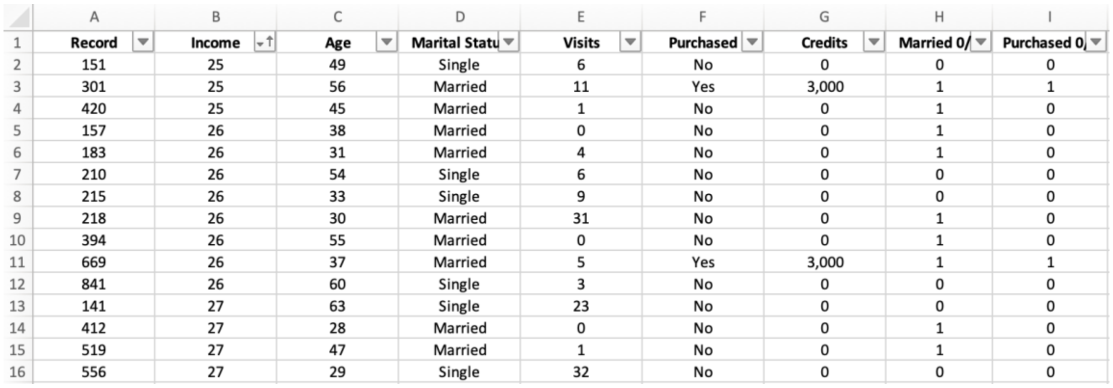
| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 852 | **Correlation** | Income | Age | Married 0/1 | Visits | Purchased 0/1 |
| 853 | Income | 1 | | | | |
| 854 | Age | 0.079 | 1 | | | |
| 855 | Married 0/1 | 0.035 | 0.140 | 1 | | |
| 856 | Visits | 0.112 | -0.015 | -0.028 | 1 | |
| 857 | Purchased 0/1 | 0.128 | -0.002 | -0.009 | 0.146 | 1 |

*f.* Income sorted highest to lowest. Purchasing credits (sometimes many) is common in the high-income group.



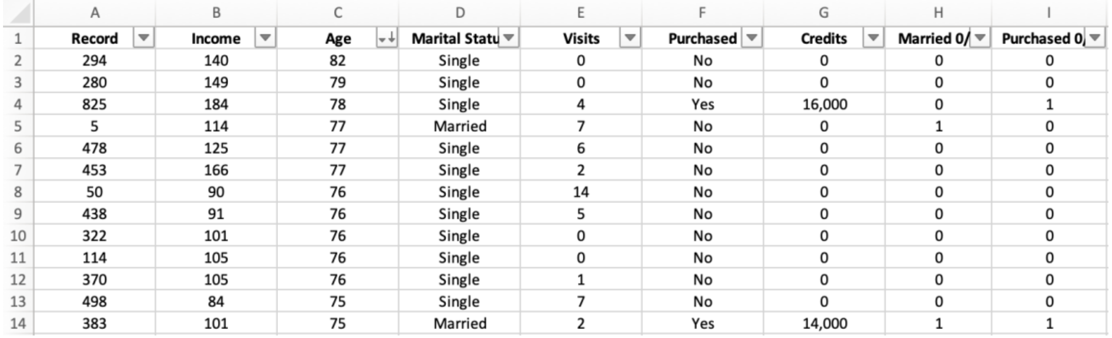
| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | **Record** | **Income** | **Age** | **Marital Statu** | **Visits** | **Purchased** | **Credits** | **Married 0/** | **Purchased 0,** |
| 2 | 357 | 999 | 30 | Married | 0 | Yes | 3,000 | 1 | 1 |
| 3 | 835 | 994 | 43 | Married | 69 | No | 0 | 1 | 0 |
| 4 | 28 | 826 | 31 | Married | 5 | Yes | 17,000 | 1 | 1 |
| 5 | 209 | 817 | 36 | Single | 5 | No | 0 | 0 | 0 |
| 6 | 348 | 808 | 34 | Single | 4 | Yes | 17,000 | 0 | 1 |
| 7 | 424 | 764 | 33 | Married | 2 | No | 0 | 1 | 0 |
| 8 | 62 | 715 | 57 | Married | 31 | Yes | 17,000 | 1 | 1 |
| 9 | 470 | 483 | 60 | Single | 0 | Yes | 3,000 | 0 | 1 |
| 10 | 115 | 471 | 39 | Single | 0 | Yes | 3,000 | 0 | 1 |
| 11 | 766 | 464 | 53 | Married | 20 | No | 0 | 1 | 0 |
| 12 | 44 | 435 | 48 | Married | 30 | Yes | 3,000 | 1 | 1 |
| 13 | 239 | 435 | 69 | Single | 4 | No | 0 | 0 | 0 |
| 14 | 723 | 417 | 56 | Married | 3 | Yes | 12,000 | 1 | 1 |

Income sorted from lowest to highest. Purchasing credits is rare among the low-income group, and then usually the minimum (3000) are purchased.



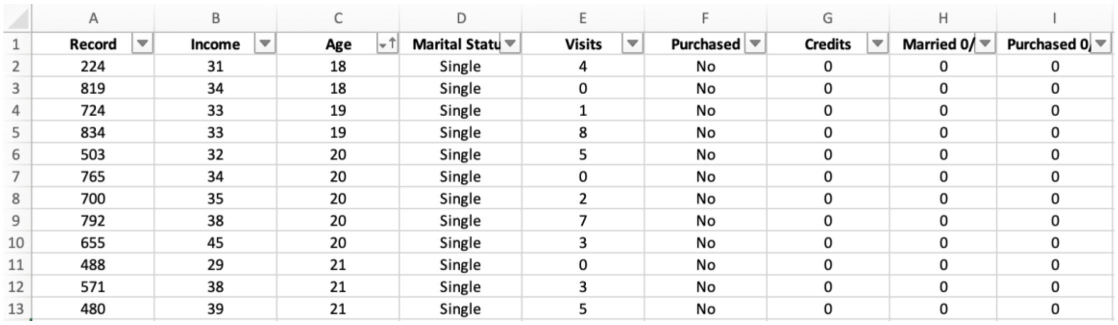
| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | **Record** | **Income** | **Age** | **Marital Statu** | **Visits** | **Purchased** | **Credits** | **Married 0/** | **Purchased 0,** |
| 2 | 151 | 25 | 49 | Single | 6 | No | 0 | 0 | 0 |
| 3 | 301 | 25 | 56 | Married | 11 | Yes | 3,000 | 1 | 1 |
| 4 | 420 | 25 | 45 | Married | 1 | No | 0 | 1 | 0 |
| 5 | 157 | 26 | 38 | Married | 0 | No | 0 | 1 | 0 |
| 6 | 183 | 26 | 31 | Married | 4 | No | 0 | 1 | 0 |
| 7 | 210 | 26 | 54 | Single | 6 | No | 0 | 0 | 0 |
| 8 | 215 | 26 | 33 | Single | 9 | No | 0 | 0 | 0 |
| 9 | 218 | 26 | 30 | Married | 31 | No | 0 | 1 | 0 |
| 10 | 394 | 26 | 55 | Married | 0 | No | 0 | 1 | 0 |
| 11 | 669 | 26 | 37 | Married | 5 | Yes | 3,000 | 1 | 1 |
| 12 | 841 | 26 | 60 | Single | 3 | No | 0 | 0 | 0 |
| 13 | 141 | 27 | 63 | Single | 23 | No | 0 | 0 | 0 |
| 14 | 412 | 27 | 28 | Married | 0 | No | 0 | 1 | 0 |
| 15 | 519 | 27 | 47 | Married | 1 | No | 0 | 1 | 0 |
| 16 | 556 | 27 | 29 | Single | 32 | No | 0 | 0 | 0 |

Age sorted highest to lowest. Purchasing credits is rare in the highest age group, but when purchased, a large number may be purchased.



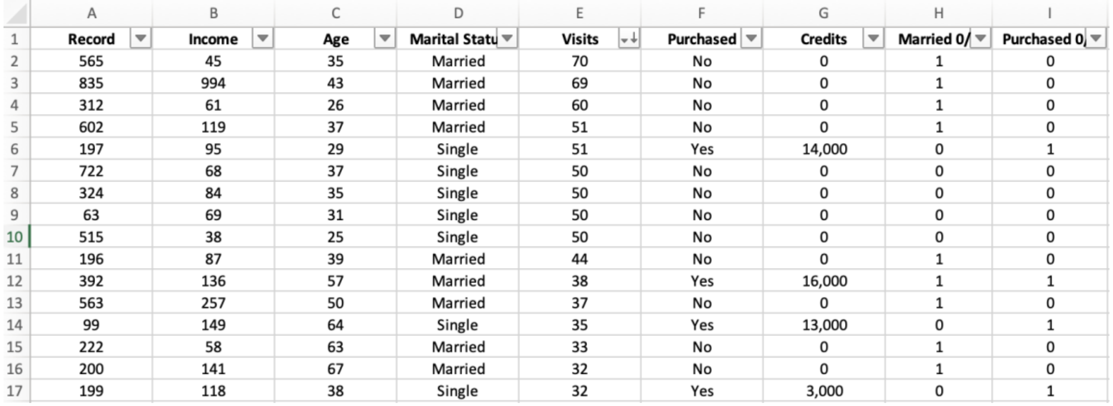
| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | **Record** | **Income** | **Age** | **Marital Statu** | **Visits** | **Purchased** | **Credits** | **Married 0/** | **Purchased 0,** |
| 2 | 294 | 140 | 82 | Single | 0 | No | 0 | 0 | 0 |
| 3 | 280 | 149 | 79 | Single | 0 | No | 0 | 0 | 0 |
| 4 | 825 | 184 | 78 | Single | 4 | Yes | 16,000 | 0 | 1 |
| 5 | 5 | 114 | 77 | Married | 7 | No | 0 | 1 | 0 |
| 6 | 478 | 125 | 77 | Single | 6 | No | 0 | 0 | 0 |
| 7 | 453 | 166 | 77 | Single | 2 | No | 0 | 0 | 0 |
| 8 | 50 | 90 | 76 | Single | 14 | No | 0 | 0 | 0 |
| 9 | 438 | 91 | 76 | Single | 5 | No | 0 | 0 | 0 |
| 10 | 322 | 101 | 76 | Single | 0 | No | 0 | 0 | 0 |
| 11 | 114 | 105 | 76 | Single | 0 | No | 0 | 0 | 0 |
| 12 | 370 | 105 | 76 | Single | 1 | No | 0 | 0 | 0 |
| 13 | 498 | 84 | 75 | Single | 7 | No | 0 | 0 | 0 |
| 14 | 383 | 101 | 75 | Married | 2 | Yes | 14,000 | 1 | 1 |

Age sorted from lowest to highest. Purchasing credits is very rare (nonexistent) among the lowest age group.



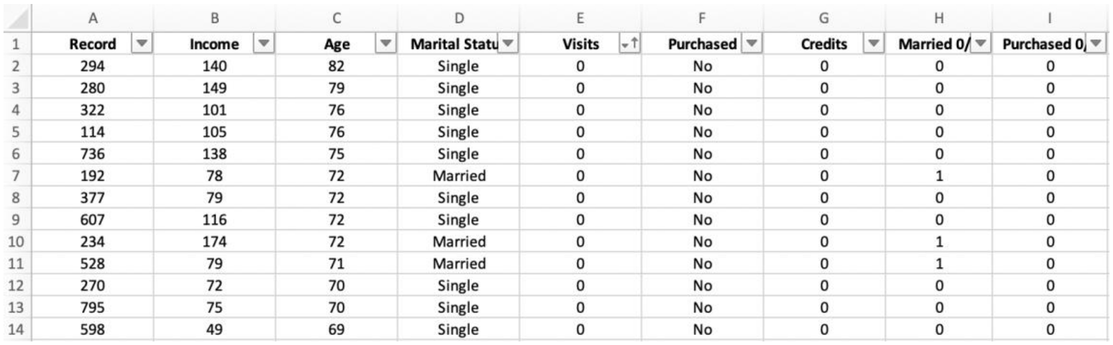
| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Record | Income | Age | Marital Statu | Visits | Purchased | Credits | Married 0/ | Purchased 0, |
| 2 | 224 | 31 | 18 | Single | 4 | No | 0 | 0 | 0 |
| 3 | 819 | 34 | 18 | Single | 0 | No | 0 | 0 | 0 |
| 4 | 724 | 33 | 19 | Single | 1 | No | 0 | 0 | 0 |
| 5 | 834 | 33 | 19 | Single | 8 | No | 0 | 0 | 0 |
| 6 | 503 | 32 | 20 | Single | 5 | No | 0 | 0 | 0 |
| 7 | 765 | 34 | 20 | Single | 0 | No | 0 | 0 | 0 |
| 8 | 700 | 35 | 20 | Single | 2 | No | 0 | 0 | 0 |
| 9 | 792 | 38 | 20 | Single | 7 | No | 0 | 0 | 0 |
| 10 | 655 | 45 | 20 | Single | 3 | No | 0 | 0 | 0 |
| 11 | 488 | 29 | 21 | Single | 0 | No | 0 | 0 | 0 |
| 12 | 571 | 38 | 21 | Single | 3 | No | 0 | 0 | 0 |
| 13 | 480 | 39 | 21 | Single | 5 | No | 0 | 0 | 0 |

Visits sorted highest to lowest. Purchasing credits is somewhat common in the highest visit group, and when purchased, a large number may be purchased.



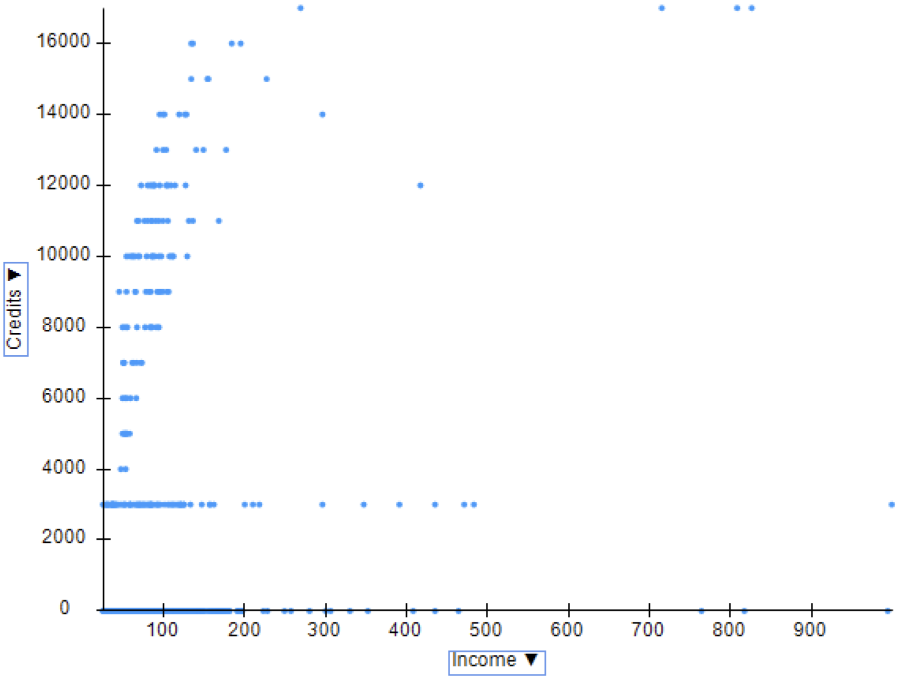
| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Record | Income | Age | Marital Statu | Visits | Purchased | Credits | Married 0/ | Purchased 0, |
| 2 | 565 | 45 | 35 | Married | 70 | No | 0 | 1 | 0 |
| 3 | 835 | 994 | 43 | Married | 69 | No | 0 | 1 | 0 |
| 4 | 312 | 61 | 26 | Married | 60 | No | 0 | 1 | 0 |
| 5 | 602 | 119 | 37 | Married | 51 | No | 0 | 1 | 0 |
| 6 | 197 | 95 | 29 | Single | 51 | Yes | 14,000 | 0 | 1 |
| 7 | 722 | 68 | 37 | Single | 50 | No | 0 | 0 | 0 |
| 8 | 324 | 84 | 35 | Single | 50 | No | 0 | 0 | 0 |
| 9 | 63 | 69 | 31 | Single | 50 | No | 0 | 0 | 0 |
| 10 | 515 | 38 | 25 | Single | 50 | No | 0 | 0 | 0 |
| 11 | 196 | 87 | 39 | Married | 44 | No | 0 | 1 | 0 |
| 12 | 392 | 136 | 57 | Married | 38 | Yes | 16,000 | 1 | 1 |
| 13 | 563 | 257 | 50 | Married | 37 | No | 0 | 1 | 0 |
| 14 | 99 | 149 | 64 | Single | 35 | Yes | 13,000 | 0 | 1 |
| 15 | 222 | 58 | 63 | Married | 33 | No | 0 | 1 | 0 |
| 16 | 200 | 141 | 67 | Married | 32 | No | 0 | 1 | 0 |
| 17 | 199 | 118 | 38 | Single | 32 | Yes | 3,000 | 0 | 1 |

Visits sorted from lowest to highest. Purchasing credits is very rare (nonexistent) among the lowest visiting group.

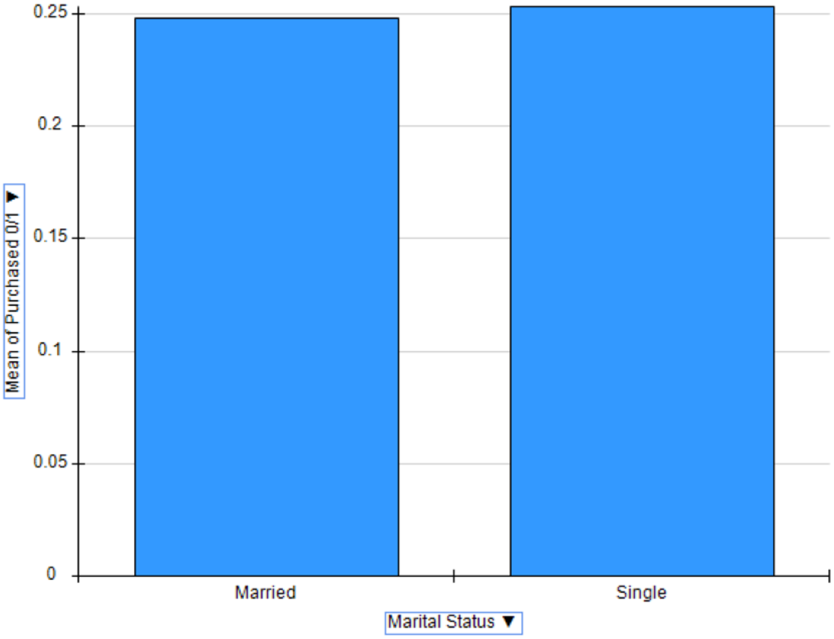


| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Record | Income | Age | Marital Statu | Visits | Purchased | Credits | Married 0/ | Purchased 0, |
| 2 | 294 | 140 | 82 | Single | 0 | No | 0 | 0 | 0 |
| 3 | 280 | 149 | 79 | Single | 0 | No | 0 | 0 | 0 |
| 4 | 322 | 101 | 76 | Single | 0 | No | 0 | 0 | 0 |
| 5 | 114 | 105 | 76 | Single | 0 | No | 0 | 0 | 0 |
| 6 | 736 | 138 | 75 | Single | 0 | No | 0 | 0 | 0 |
| 7 | 192 | 78 | 72 | Married | 0 | No | 0 | 1 | 0 |
| 8 | 377 | 79 | 72 | Single | 0 | No | 0 | 0 | 0 |
| 9 | 607 | 116 | 72 | Single | 0 | No | 0 | 0 | 0 |
| 10 | 234 | 174 | 72 | Married | 0 | No | 0 | 1 | 0 |
| 11 | 528 | 79 | 71 | Married | 0 | No | 0 | 1 | 0 |
| 12 | 270 | 72 | 70 | Single | 0 | No | 0 | 0 | 0 |
| 13 | 795 | 75 | 70 | Single | 0 | No | 0 | 0 | 0 |
| 14 | 598 | 49 | 69 | Single | 0 | No | 0 | 0 | 0 |

*g.* There appears to be a positive correlation between credits and income.

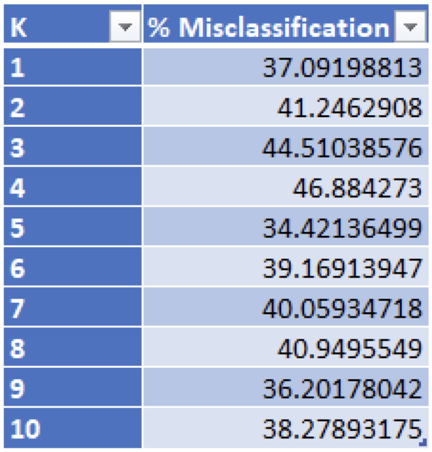


*h.* Married or single does not appear to change the probability of purchase.



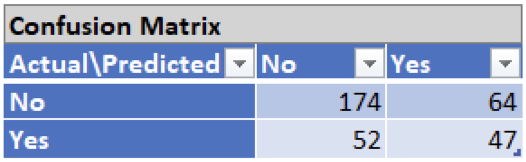
*i.* Based on the correlation table from part *e*, it appears *Income* and *Visits* are most closely correlated with whether a purchase is made. The other predictor variables show little correlation.

*j.* The KNN algorithm performs best with *k* = 5.



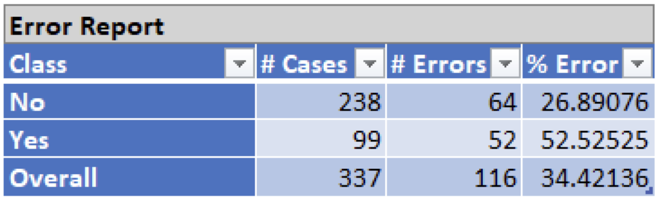
| K | % Misclassification |
|---|---|
| 1 | 37.09198813 |
| 2 | 41.2462908 |
| 3 | 44.51038576 |
| 4 | 46.884273 |
| 5 | 34.42136499 |
| 6 | 39.16913947 |
| 7 | 40.05934718 |
| 8 | 40.9495549 |
| 9 | 36.20178042 |
| 10 | 38.27893175 |

*k.* Using *k* = 5 for KNN, Accuracy = 65.6%, Specificity = 73.1%, Sensitivity = 47.4%.



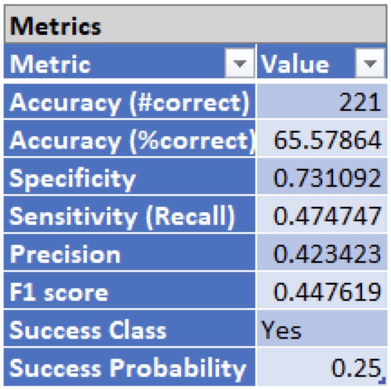
Confusion Matrix

| Actual\Predicted | No | Yes |
|---|---|---|
| No | 174 | 64 |
| Yes | 52 | 47 |



Error Report

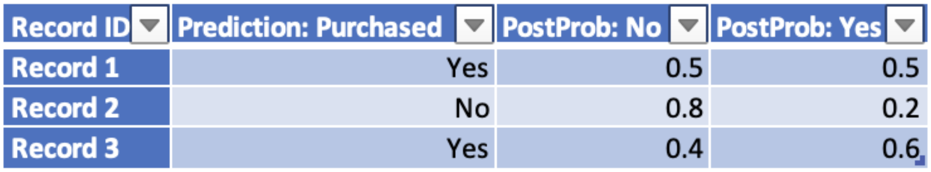
| Class | # Cases | # Errors | % Error |
|---|---|---|---|
| No | 238 | 64 | 26.89076 |
| Yes | 99 | 52 | 52.52525 |
| Overall | 337 | 116 | 34.42136 |



Metrics

| Metric | Value |
|---|---|
| Accuracy (#correct) | 221 |
| Accuracy (%correct) | 65.57864 |
| Specificity | 0.731092 |
| Sensitivity (Recall) | 0.474747 |
| Precision | 0.423423 |
| F1 score | 0.447619 |
| Success Class | Yes |
| Success Probability | 0.25 |

*l.* Using *Income* and *Visits* as predictor variables and *k* = 5, the first and third potential buyers are classified as likely to purchase. The estimated probability of purchase for each of the three buyers is 0.5, 0.2, and 0.6, respectively.



| Record ID | Prediction: Purchased | PostProb: No | PostProb: Yes |
|---|---|---|---|
| Record 1 | Yes | 0.5 | 0.5 |
| Record 2 | No | 0.8 | 0.2 |
| Record 3 | Yes | 0.4 | 0.6 |