Logistic Regression for the Classification of Credit Risk

Sammie Haskin, 2nd Year MSAS student

KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING
School of Data Science and Analytics

Under the Supervision of Dr. Jennifer Priestley, Professor, School of Data Science and Analytics

Introduction

The utilization of statistics, if performed effectively, can be the deciding factor as to whether a credit lender is successful. While the approval of credit for a consumer who will pay on time results in $250 in revenue on average, the approval of credit for a consumer who later defaults can be a particularly costly mistake leading to approximately $750 in losses on average per instance. As a result, the successful identification of consumers who will not default is a priority for credit lenders.

With this in mind, we sought to create a model that could be utilized to effectively predict whether consumers are likely to default. The data was gathered by from a major credit bureau and contained 1,462,955 observations of consumers prior to approval. Another data set contained the 17,244,104 observations for the longitudinal post-hoc performance of each consumer after approval.

Utilizing over a million observations of consumers, over 300 attributes per consumer, and a logistic regression procedure, we sought to create an effective model utilizing only 10 attributes for the purpose of the classification of the risk of defaulting for each consumer.

Methods

In this analysis, credit defaulting was defined as any scenario in which a consumer was more than two cycles late on their credit payment. Attributes that were known prior to the approval of consumers were to be further examined for suitability in the modeling phase. After the removal of observations in which the majority of attributes were missing or no credit history for consumers was recorded, 1,255,429 observations were available for model creation and evaluation. A **stratified median imputation** procedure was utilized for attributes where less than 30% of the total observations were missing. In this step, 124 potential predictors were retained. A **variable clustering** procedure was subsequently implemented for the identification of 20 attributes that could each uniquely account for the some of the variation of the probability of defaulting among consumers.

Through the subsequent implementation of the techniques of **attribute assessment**, **discretization**, and **transformation**, attributes that could contribute to the classification of credit risk were identified and prepared. Using these 18 attributes, a **logistic regression** procedure was chosen for modeling, because the technique allows for the evaluation of the probability of default and the evaluation of the expected effect of each credit attribute on the resulting probability of defaulting. To implement the logistic regression model, the data was then randomly split into 80% and 20% training and testing splits, respectively. Utilizing the **information value** statistic, a **backward selection** method, and other methods of attribute evaluation, the number of predictors selected was reduced to 10 attributes for the purpose of creating a less complex, interpretable model that would simultaneously maintain a high degree of effectiveness.

Utilizing well known methods in model evaluation as well as domain knowledge as to the cost of correct and incorrect classifications, the practical implications of the model's use were evaluated and an optimal cut point for the predicted probability of defaulting was identified such that the revenue from credit lending could be increased. Consumers with probabilities of defaulting below this cut point were to be granted credit. From these and other statistics, the suitability of the model for use was evaluated.

Figure 1: ROC Curve in the Classification of Credit Risk

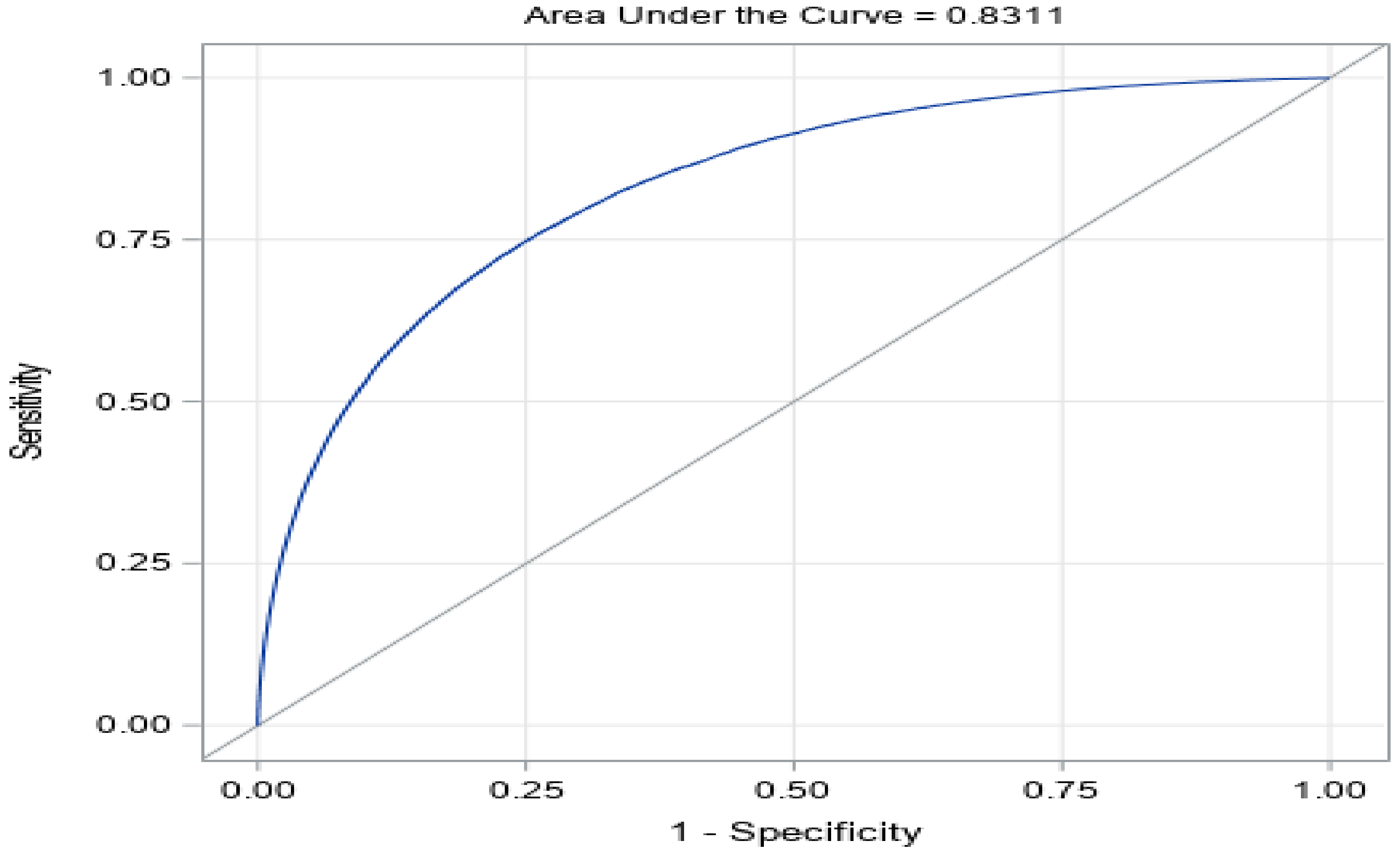

Figure 2: Expected Profits by Cut Point for the Classification of Credit Risk

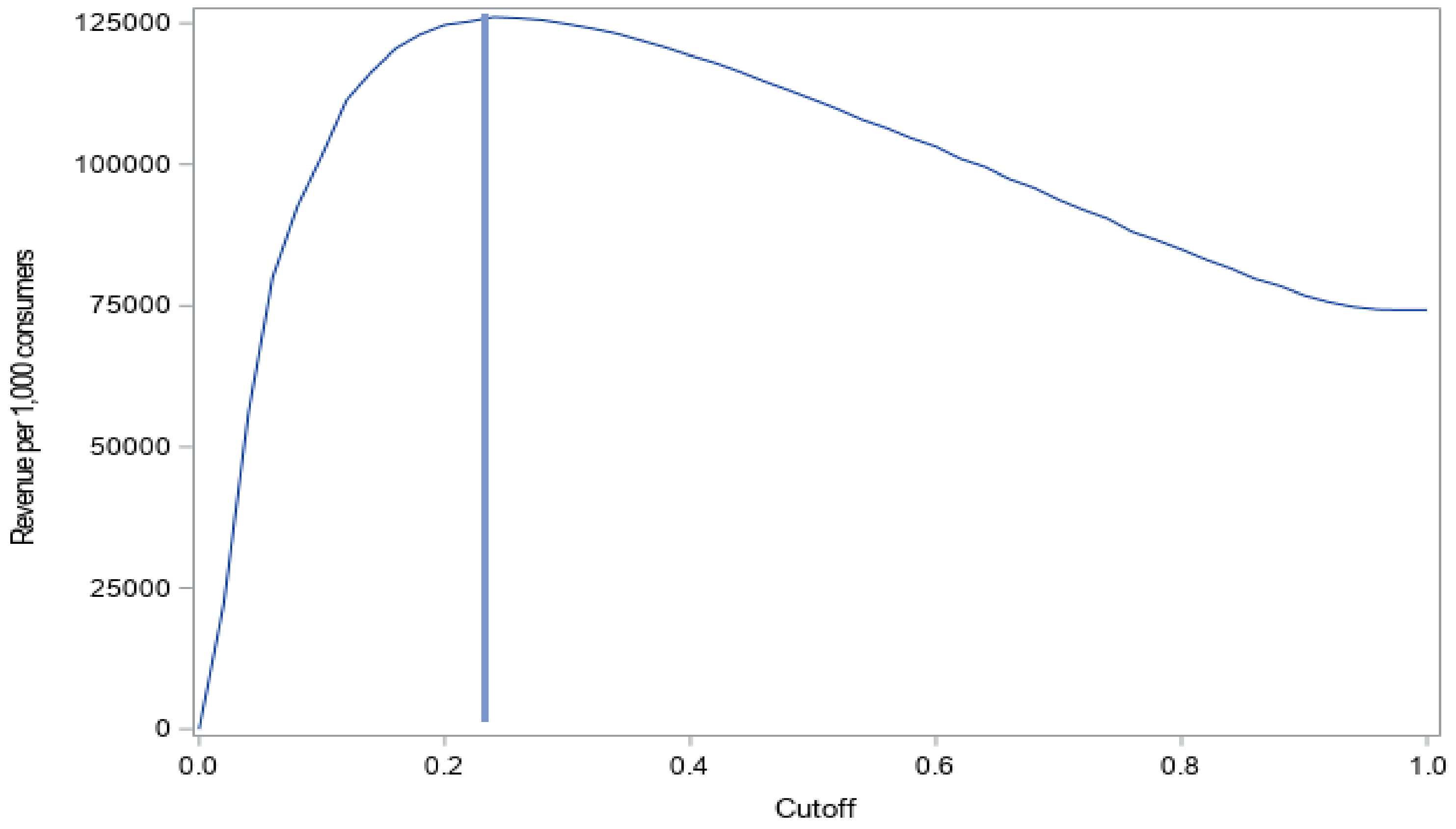

Table 1: Confusion Matrix at the Established Cut Point of 25%

	Actual Class	
Predicted Class	Did Not Default	Defaulted
Low Risk	177225 (70.58)	16807 (6.69%)
High Risk	29740 (11.84%)	27313 (10.88%)

Table 2: Calculations for the KS Statistic and Lift of the Logistic Regression Model

Bin	N	N Good Credit	N Bad Credit	% Good	% Bad	Cumulative % Good Credit	Cumulative % Bad Credit	KS	Lift
1	25109	8775	16334	0.042	0.370	4.24	37.02	32.78	3.70
2	25109	16063	9046	0.078	0.205	12.00	57.53	45.52	2.88
3	25109	19227	5882	0.093	0.133	21.29	70.86	49.57	2.36
4	25109	20840	4269	0.101	0.097	31.36	80.53	49.17	2.01
5	25109	22211	2898	0.107	0.066	42.09	87.10	45.01	1.74
6	25109	22643	2466	0.109	0.056	53.03	92.69	39.66	1.54
7	25109	23538	1571	0.114	0.036	64.41	96.25	31.85	1.38
8	25109	24201	908	0.117	0.021	76.10	98.31	22.21	1.23
9	25109	24635	474	0.119	0.011	88.00	99.38	11.38	1.10
10	25104	24832	272	0.120	0.006	100.00	100.00	0.00	1.00

Results

The initial logistic regression model contained 18 attributes and had a concordance statistic of .831 on the training data. Iteratively removing attributes with the smallest chi squared test statistics, the model was reduced to 10 predictors yet maintained a **c statistic of .831**. Observing the Receiving Operating Characteristic curve, the produced model functioned with an effectiveness that was significantly greater than what could be expected by chance alone.

In the evaluation of the expected profitability of the model using the validation data set, a curve of the revenues expected at each possible cut point for the probability of defaulting was produced. Establishing the cutoff for the predicted probability of defaulting as .25, the approval of consumers for credit who are less than 25% likely to default was calculated to **net $126,256.05 per 1000 applicants**. At this cutoff, the **precision** of the identification of consumers that would not default was **.758** while the **recall** out of all consumers that would not default was **.9134**. With an **F1 score of .796**, the model indicated a high degree of accuracy in the process of determining consumers who would not default.

Further evaluating the effectiveness of the model, the Kolmogorov–Smirnov (KS) and Lift statistics were calculated on the testing data and are shown in Table 2. As indicated by the **KS statistic of 49.56%**, the model performed effectively in the prediction of the risk of defaulting for consumers that did and did not default. As evaluated with the **Lift statistic**, it was determined that the model could correctly identify **71%** of the consumers that defaulted from only **30%** of the data. These statistics indicate that an effective model was created that could aid in the evaluation of the risk of defaulting per consumer given their attributes known beforehand.

Conclusion

Utilizing several statistical procedures in the preparation of the data such as the defining of credit risk, the imputation of missing values, the variable clustering process for attribute selection, and the discretization of potentially useful attributes, the foundational steps toward the classification of credit risk were implemented in a manner that insured success in the model building process. With each of these foundational steps in data preparation, an effective model was created that could be successfully implemented in the evaluation of the suitability of credit approval for subprime consumers.

The use of such a model in credit risk classification is far-reaching in that it allows for a less complex method of determining whether or not to grant credit to a given consumer. The additional benefit of the logistic regression model for credit scoring in this circumstance is that it also allows for relatively intuitive interpretations of why consumers were not granted credit given the nature of the procedure and the relatively small number of predictors. With successive iterations in the modification of the data preparation stage, it is possible that the created model could be improved to further increase the profitability in the approval of credit for subprime consumers.

References

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences, 2nd Edition. Routledge

Lin, A. Z. (2013). Variable Reduction in SAS by Using Weight of Evidence and Information Value. SAS Global Forum.

Siddiqi, N. (2005). Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring. John Wiley & Sons, Inc.