Bookbinders Case Study

Collin Real, Leonel Salazar, Seth Harris, Joaquin Ramirez

# I. Executive Summary

In this case study, we explored three types of models to predict the customers most likely to purchase The Art of History of Florence for BBBC’s marketing campaign: linear regression, logistic regression, and support vector machines (SVMs). BBBC wants to develop a highly accurate model to more effectively market to target customers, potentially increasing overall gross profit by increasing book revenue and decreasing mailing costs by only advertising to customers most likely to purchase their products. The logistic regression model provided the highest prediction accuracy, AUC, and sensitivity for purchased books.

# II. The Problem

BBBC’s main challenge is developing an effective marketing strategy that reduces mailing costs by identifying target customers using historical purchasing data and customer characteristics to predict customer purchases. Promotional advertisements will only be mailed to this key demographic rather than every customer, saving on costs while potentially growing sales and gross profit. One issue the company faces is the class imbalance in predicting customer purchases: customers who did not purchase the book is significantly higher those those who did which can potentially skew the model’s performance. Additionally, BBBC must compare the trade-offs between models that are highly accurate (i.e., SVMs) and more interpretable models (i.e., logistic regression).

# III. Review of Related Literature

## Linear Regression

* Dependent variable is **continuous**.
* Estimates the relationship between predictors and the dependent variable by fitting a linear equation.
* Easy implementation and interpretation.
* Model assumptions:
  + **Linear relationship** between the predictors and the dependent variable.
  + Residual errors are **independent** (no autocorrelation) of each other.
  + **Multicollinearity** - independent variables are not perfectly correlated.
  + **Homoscedasticity** - error variance is **constant** across all levels of the independent variables.
  + Residuals are **normally distributed**.

**The linear regression model is not appropriate for our Bookbinder’s case study because our models are predicting a binary (i.e., purchase vs. non-purchase), not a continuous outcome. Since the linear regression can predict values outside of the [0,1] range, probability cannot be estimated correctly. Homoscedasticity, linearity, and normality assumptions are also violated. This model is only included in the case study for comparison to other models.**

## Logistic Regression

* Binary classification problems (i.e., purchase vs. no purchase)
* Estimates the probablity of an event occurring by modeling the relationship between 1+ predictors and a binary outcome.
* Easy implementation and interpretation.
* Model assumptions:
  + **Dependent variable** is a binary outcome.
  + **Linear relationship** between the predictors and log-odds of the dependent variable; however, this assumption is not always true which weakens the effectiveness of this model. **Non-linear** relationship between the predictors and dependent variable.
  + **Independent observations**
  + **Multicollinearity** - independent variables are not perfectly correlated.
  + **Large sample size**
  + **Homoscedasticity** - error variance follows binomial distribution.

## Support Vector Machine (SVM)

* Classification problems for both linear and non-linear outcomes.
* Identifies the hyperplane with maximum separation between the two classes (i.e., purchase vs. non-purchase) in the feature space.
* Capture non-linear relationships via multiple kernels:
  + linear
  + radial basis function (RBF)
  + polynomial
  + sigmoid
* Model flexibility and effectively identifies complex relationships for non-linear problems.
* Less interpretable and difficulty selecting the correct kernel/tuning parameters.
* Model assumptions:
  + **Linearly separable classes** - Linear SVM assumption needed to find a hyperplane separating the classes. RBF or polynomial non-linear kernels are used if the assumption is not satisfied.
  + **Independent observations**
  + **Multicollinearity** - independent variables are not perfectly correlated.
  + **Feature scaling** - accounts for distances between data points. Unscaled data might affect model results.
  + **Kernels** - assumes the correct kernel is appropriately chosen.

# IV. Methodology

The primary objective of our case study was to identify highest opportunity customers to increase the efficiency of BBBC’s direct marketing efforts by reducing costs and creating potential sales opportunities.

## Data Preprocessing & Exploration

The first step of our analysis was importing our data sources and preprocessing. The data sources were two Excel spreadsheets labeled training and test. After importing both files as separate datasets, we concatenated them into one dataframe and checked for duplicates.Twenty-five were identified and removed. Combining the two data sources allows for training on more data observations, potentially increasing the overall accuracy of our final models. Additionally, building a model on a predefined training/test sets is not appropriate for building a robust model for real world applications. Using our concatenated dataframe, we applied an 70/30 training and test split which randomizes the data our models use for training and predictions. Variables Choice and Gender were converted from numeric to factor data types since they are categorical. The dataset contained zero missing values, so data imputation was not necessary.

Next, we visualized the distribution of numeric variables using histogram plots:

A graph of numbers and a number of items

Description automatically generated with medium confidence

The histograms illustrate that Amount\_purchased is the only variable whose values are normally distributed. Next, we calculated and visualized the correlation matrix using the correlation plot to identify and remove highly correlated variables from our model. Predictors First\_purchase and Last\_purchase had a significant positive correlation of **0.81**, so we removed Last\_purchase from our list of features after concluding that it also has strong positive correlations with other predictors. Observations was also removed from our features because it is a poor predictor and would affect our modeling results.

A diagram of a number of items

Description automatically generated with medium confidence

Finally, we centered and scaled our dataset. Amount\_purchased is measured at a different scale (monetarily) compared to the other numeric variables which could create bias in our model.

## Class Imbalance

The next step of our analysis was using resampling techniques to fix the class imbalance in our data. Our dependent variable Choice contains 400 customers who purchased the book and 1200 customers who did not purchase, resulting in a majority of non-purchase customers. Class imbalances can produce biased models favoring the majority class and decreased the accuracy of our models. We applied the **both** technique which created equal observations for our Purchasers class, potentially improving our models’ prediction capabilities.

## Model Selection

Linear regression, logistic regression. RBF SVM, and polynomial SVM models were explored with interaction terms to predict customer behavior. The linear regression model is used only for model comparison due to the binary nature of our dependent variable. The logistic regression model provided us with the **odds ratios** which identifies each predictors’s likelihood of influencing a customer’s purchase. Both linear and non-linear kernel SVMs were implemented for more effective pattern recognition.

## Model Evaluation

10-fold cross-validation was used to evaluate our models, acting against potential overfitting and providing insight on each model’s ability in generalizing unfamiliar data. Once training and cross-validation was completed, each model’s overall performance was evaluated to determine the most optimal for BBBC’s marketing initiative. Accuracy, recall, precision, and other metrics provided further insight on the trade-offs between each model.

# V. Data

The case report used a subset of data available to BBBC, provided as two Excel spreadsheets: one training set and one test. The training data consisted of 400 customers who purchased a book and 1200 customers who did not. The test set included 2300 data observations. The training and test files were combined into one dataframe to increase the amount of training data for our models. The dependent variable of the analysis was Choice - purchase or not purchase of the book. BBBC also selected multiple independent variables to potentially explain customer purchasing behavior:

A white page with black text

Description automatically generated

# VI. Findings (Results)

Our findings illustrate that the **Logistic Regression** model had the highest accuracy rate equaling **87.01%**. The C hyperparameter in SVMs are used to control the error or margin of the model: lower C means lower error. The Gamma hyperparameter chooses the amount of curvature to apply to the decision boundary: high Gamma means higher curvature. Linear Regression had the second highest accuracy, followed by the RBF SVM, and Polynomial SVM. The RBF SVM produced the highest Area Under the Curve (AUC), indicating that it had the overall best performance among its competitors.

## ROC Curve and AUC

A graph of a model

Description automatically generated

## Model Performances

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **AUC** | **Accuracy** | **Precision** | **Recall** | **F1\_score** |
| Logistic Regression | 0.817 | 0.870 | 0.702 | 0.308 | 0.428 |
| Linear Regression | 0.815 | 0.842 | 0.000 | 0.000 | 0.000 |
| RBF SVM | 0.849 | 0.795 | 0.416 | 0.738 | 0.532 |
| Polynomial SVM | 0.838 | 0.788 | 0.404 | 0.731 | 0.521 |

## Logistic Regression

Our first logistic regression model included the features from our resampled dataset (excluding Observation and Last\_purchase) as well as every interaction term. After identifying the significant predictors/interaction terms from our initial model, we built a second logistic regression that only included significant features from the first model. We continued this process until all predictors in our final model were statistically significant. The final model predictors: **GenderMale**, **Frequency**, **First\_purchase** , **P\_Child**, **P\_Cook**, **P\_DIY**, **P\_Art**, and **Amount\_purchased.** These predictors were included as features in the rest of our models.

### Model Summary

A screenshot of a computer program

Description automatically generated

### Model Summary

A screen shot of a computer screen

Description automatically generated

### 

### Checking Assumptions

Our logistic regression passes the binary outcome, independent observations, and adequate sample size assumptions. To check for multicollinearity, we use the variance inflation factor (VIF) to identify any values greater than 5. To check for homoscedasticity, diagnostic plots were generated for the model.

### VIF

A black rectangular object with white lines

Description automatically generated

**First\_purchase** VIF value is greater than 5, but we did not remove it from out model since its correlation to other independent variables does not raise any concern.

### Homoscedasticity

A collage of graphs

Description automatically generated

### Variable Importance

A graph with blue bars

Description automatically generated

## Radial Basis Function SVM - Model Application

To identify our target customers, we inspected the **odds ratios** calculated from our logistic regression model. The best predictor for predicting customer purchases is the number of art books purchased during their visit. Customers who purchased an art book are 118% more likely buy a Art History of Florence from BBBC. Another strong indicator of customer engagement is the number of months since their first visit (i.e., the longer they’ve shopped at BBBC, they are 74% more likely to buy). The strength of this predictor demonstrates customer loyalty to Bookbinders. The Frequency predictor supports this claim: infrequent or one-time shoppers are 64% less likely to purchase because their attention is elsewhere. The amount spent at Bookbinders also points towards potential targets, and further complements our customer loyalty claim. Finally, women are more likely to purchase the book.

To calculate the potential gross profit generation, we used the compare matrix values from our radial basis SVM model. The confusion matrix values: - true positive: 689 - false negative: 104 - false positive: 5 - false negative: 26

The resulting gross profit calculations was **$8,271** via targeted promotional advertisements and **$47,961** for mailing the whole list. These results demonstrate a potential opportunity cost that arises in customer segmentation where models fail to identify target customers who could have been a potential sale. Despite the theoretical mail costs savings by narrowing our “target” audience, our model produced significantly more false negatives and overlooked customers with high potential.

## Conclusion & Recommendations

Our group recommended using the Radial Basis Function SVM model to predict customer purchases, a binary outcome. We also focused heavily on the logistic regression model, another binary classification model, to gain a better understanding of the features’ affect on its model. Our results show that, despite producing models with high accuracy, predicting outcomes in binary classification problems is tedious and inherently risky. We used advantages of the logit model, such as investigating the coefficients, various interaction terms, and their overall performance. Although we discovered various significant predictors, our confusion matrix would predict a disappointing amount of false positives and negatives. These false positives and negatives sabotage BBBC whose marketing campaign is guided by our model’s recommendations. On the contrary, our SVM models generated less false positives and negatives, but interpreting the model to understand the most effective predictors is not as simple. Despite both models solving the same binary problem, it’s also ineffective to use the advantages of one binary classification model to draw a conclusion and assume it’s true for another binary model.

Despite these advantage and disadvantages, companies should always develop knowledge and expertise to understand their data and how to use it effectively. Companies gain a competitive advantage by iterating over and experimenting with various models, even if they aren’t deploying them in their business operations. Our recommendation to BBBC is to provide more data for niche modeling exercises like the art history book, especially for binary classification problems. Humans can be complex, emotional, irrational, generous, and kind over a couple of hours time, so making generalizations about there customers based off a few thousand observations and less than 10 predictors will not be super effective. Gathering customer data and sending out automated survey emails can begin this process. From a technical standpoint, building their own Python package for their specific modeling needs will help streamline their data science process. The company should develop and maintain this package through a shared GitLab repository, constantly pulling updates from others and pushing their changes. Within the package, their classes and functions should be written in a modular way to prevent their tools from becoming unmaintainable. The team should create .bat files to automate any computer administration tasks, and schedule them to execute at their desired update cadence.

Our final recommendation is for BBBC to retire its mail delivery operations and transition completely to the online marketplace. The technical skills we identified previously will facilitate this transition into the online world which offers greater opportunity for customer engagement, influence, and purchases. Their online operations should include data collection on competitors, customers, and major events they can incorporate into their products or services. It is critical for BBBC to implement efficient technologies to easily store, retrieve, and backup data essential to its model building and e-commerce operations. BBBC’s transition into cyberspace is not only crucial for its model building, but for its overall survival.

Appendix

## Radial SVM

### Summary

A screenshot of a computer

Description automatically generated

### Best Parameters

A black rectangular object with white lines

Description automatically generated

### Confusion Matrix

A screenshot of a computer screen

Description automatically generated

## Polynomial SVM

### Summary

A screenshot of a computer program

Description automatically generated

### Best Parameters

A black rectangular object with white lines

Description automatically generated

### Confusion Matrix

A screen shot of a black screen

Description automatically generated

## Linear Regression

### Summary

A screenshot of a computer screen

Description automatically generated

### Confusion Matrix

A screen shot of a computer screen

Description automatically generated