Direct Marketing

Predictive Analytics

Bellevue University

Fall 2020

Torrey Capobianco

Conrad Ibanez

Edris Safari

Executive Summary

Predictive analytics can be used to improve the effectiveness of direct mailing by identifying which customers will buy and forecasting how much they will spend. This report provides an analysis of customer data and attempts to predict the best customers to target direct mailing catalogs for maximizing profits. A data science project was completed using the cross-industry standard process for data mining, also known as CRISP-DM. The major phases of the process include business understanding, data understanding, data preparation, modeling, and evaluation. Deployment is pending upon additional analysis.

The company wishes to know which households are predicted to spend more from receiving a catalog, to increase its return on investment in direct mailing. This project analyzes a company's current customer data consisting of 1000 samples and ten attributes. We classify customers into three groups based on the amount spent: Low, Medium, and High, and attempt to identify relationships between customer attributes and spending.

Predictive analytics was completed using 80% of the data as the training set and the remaining 20% as the test set. Five different models were created for predicting the type of spender. They include K-nearest neighbors (KNN), linear regression, logistic regression, decision tree, and random forest. Random forest and decision tree outperformed the other models by far with an accuracy of 82% and 81% respectively.

The findings of this project suggest that a direct mailing campaign can be deployed using random forest to target a select group of customers that are high spenders. However, due to the limited amount of data, our recommendation would be to gather additional customer information, especially samples, and rebuild the models. Additional customer features may uncover other strong relationships to spending and help improve accuracy of the predictive models.

Business Understanding

Background

Marketing is a business sector that can leverage predictive analytics to determine the type of customer to send advertisements. In their paper, John Blackwell and Tracy DeCanio describe how The Nature Conservancy in Arlington, VA, uses predictive analytics to determine recipients for their fundraising mail solicitation "based on which donors are the most likely to respond with the largest gifts." Our project looks at a company that conducts their business sales using direct mailers of catalogs sent to potential customers. Unknowing of who is more likely to purchase an item over another person, catalogs are mailed out to homes in the hopes that a person will order from their store. A cold calling of mailers can hinder the growth of the company due to the wasted resources from the ineffective marketing. Households can receive catalogs that are irrelevant to their needs or received by those who will make significant purchases.

Problem Statement

There are three types of data that can be used to build a predictive model for direct mailing. They include data on actual customer responses to mailing campaigns, data on customer purchases over a past period, and data indicating customer interest in products (Breur and Paas). The company has collected data on previous customers who received catalogs and made purchases. Company executives would like to use the information to determine which households to send their mailers in order to increase return on investment. The problem this project addresses is to determine what type of spender a customer would be based on the given attributes in the data. They can then focus their efforts on expanding their regions to customers who are likely to spend more. Through predictive analytics, we present multiple models to identify how much a person is predicted to spend and what type of spender we would label them

as: a low, medium, or high spender. A recommendation of which model the company should use will be given based on model accuracy.

Data Understanding

The data used in this project was retrieved from Kaggle. It contains ten features and 1,000 samples. The amount spent feature is our target variable, which is the value that we are trying to predict. The remaining nine features provided are age, gender, type of home (rent or own), marital status, location, salary, children, history with the company, and catalogs. Salary, children, catalogs, and amount spent variables are continuous variables. All remaining variables are categorical. Table 1 describes the values for each categorical variable.

Variable	Value
Age	Young, Middle, Old
Gender	Male, Female
OwnHome	Own, Rent
Married	Single, Married
Location	Far, Close
History	Low, Medium, High

Table 1: Data dictionary

Exploratory Data Analysis

The amount spent is the dependent variable in our exploratory analysis while the other variables are evaluated as independent variables. For Gender, OwnHome, and Married attributes, the data is split almost 50% for each attribute value. For Age, about 51% are middle age, 29% young, and 20% old. However, the numerical age ranges for classifications are unknown.

Additionally, there were questions around the other variables such as History, Catalogs, and Location, although some assumptions can be made. For example, History may reflect the number of purchases or length of time an individual has been a customer. The Catalogs attribute is likely the number of catalogs that have been mailed out to the customer, and Location may be the distance, whether close or far, that the customer is from a physical store. Catalogs were also

evenly dispersed with values of 6, 12, 18, and 24. This dataset did not contain any households that received a catalog but did not spend any money.

We created various graphs as part of our EDA including a correlation heatmap, histograms, boxplots, and scatterplots with regression lines. In Table 2, the correlation heatmap indicates a strong positive relationship between the amount spent and salary. The heatmap also shows a positive relationship between amount spent and number of catalogs, which is also reflected in the scatterplot in Figure 1.

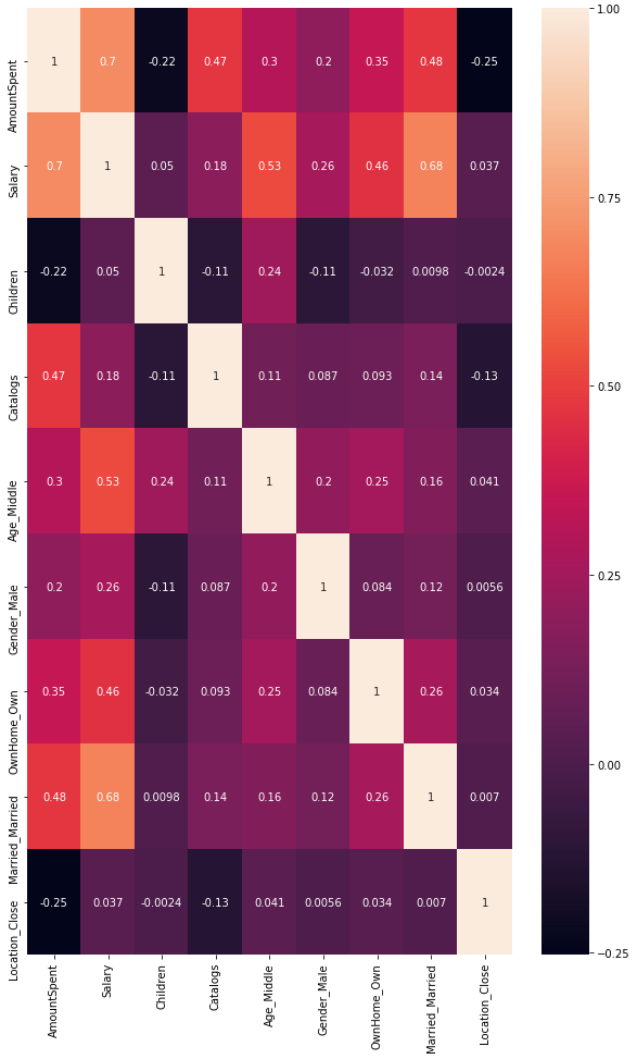

Table 2: Correlation heatmap

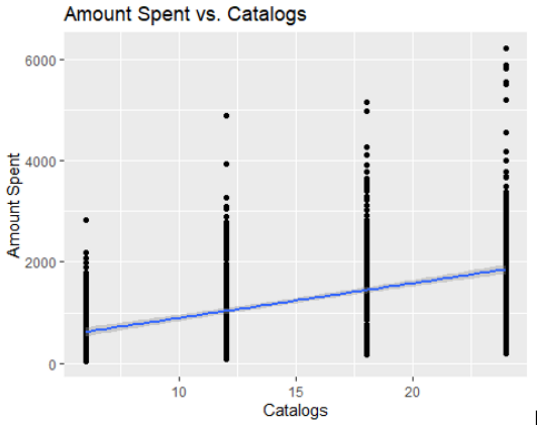

Figure 2: Amount spent vs. catalogs with regression line

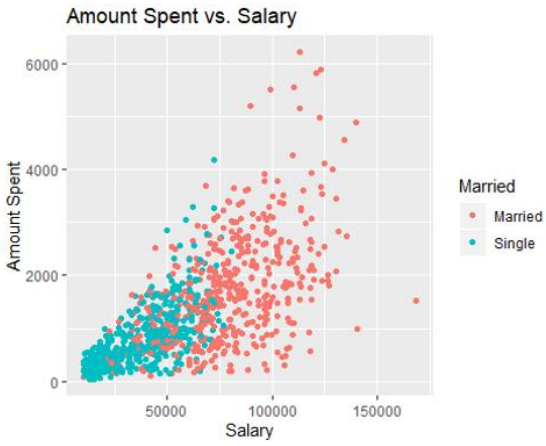

Figure 1: Amount spent vs. salary with married

The heatmap also shows a strong correlation between those who are married and salary and can be seen in the scatterplot for amount spent vs. salary along with the married status as labels in Figure 2. There is also a positive relationship between salary and those who own homes and are middle-aged. We observe a negative relationship between amount spent and the number of children. Our exploratory data analysis indicates people who are married and middle-aged, own a home, and have less children may spend more.

Data Preparation

To prepare the data for modeling, we needed to classify our spenders and look at handling categorical data and missing values. The data was cleaned using the R programing language. The only column with missing values was History with 303 values missing. We were unable to determine this value based on traditional means to fill in missing values. Due to the low number of samples in the dataset, we decided to drop the history variable from the data.

Next, we one hot encoded the categorical variables and dropped one of the exploded variables to avoid dummy trap. The columns were renamed accordingly. For example, the exploded variable ‘married_single’ was renamed to ‘married’; where the value of 1 means married, and 0 single. The resulting dataset was written to a csv file for the consumption of the linear regression modeling program.

Secondly, we created a cleaned dataset for our classification models. The data was prepped the same as before, but included all exploded dummy variables. We grouped each sample into a type of spender classification based on the amount spent. The labels were low, medium, and high. These groups were dependent upon the quantiles for amount spent. Table 3 details how the spending class was derived. The amount spent column was then dropped as our target variable was now the class.

Class	Range	Amount
Low	Below Median Quantile	≤ $962.00
Medium	Between Median and 3rd	> $962.00 & ≤ $1688.50
High	Above 3rd Quantile	> $1688.50

Table 3: Classification of spender type

The resulting dataset was written to a csv file for the consumption of the classification modeling programs. The K-nearest neighbors model used this cleaned version, but we normalized the continuous variables so that they were all on the same scale. This model also dropped one of every categorical variable to avoid multicollinearity.

Methods

We present five predictive modeling methods: K-nearest neighbors, decision tree, linear regression, logistic regression, and random forest. 80% of the data was used to train the models. To understand the model accuracy with unseen data, the remaining 20% of the data was reserved for the testing set. The same training and testing set was used across all models for consistency.

K-Nearest Neighbors

K-nearest neighbors (KNN) is a classification model used in predictive analytics to predict which type of spender a person would be, a low, medium, or high spender. The KNN classifier model is designed to predict which group the person belongs to by looking at the person's nearest neighbor's class, meaning they would have similar spending behavior. To find the optimal values K, the number of nearest neighbors to pick, the training and testing sets were cycled through 10 values for K neighbors. The value of K where the training and testing set yielded the highest accuracy was the value 2 nearest neighbors. Scikit-learn's KNeighborsClassifier was the tool used for this model.

Using 2 nearest neighbors, the testing set resulted in 71.5% accuracy. The precision score looks at the proportion of predictions predicted correctly for each class. Using the precision

score, KNN did well with predicting low spenders, with 80% predicted correctly. It did not do well with predicting high or medium spenders, with 64.62% and 52% respectively.

Decision Tree

Decision tree is another model that can be used when the target is continuous or categorical, and the predictors can be a mix of categorical and continuous variables. In the decision tree modeling, the entire dataset is placed in a root node. Then the algorithm splits the node to child nodes and proceeds to split child nodes until their respective leaf nodes are reached. The measure and strength of the splits are evaluated and optimized by the decision tree algorithm. The Gini algorithm was used because the target variable is categorical and the predictors are mixed.

Scikit-learn's DecisionTreeClassifier was used to train and test the dataset. Its max_depth parameter controls the depth of the decision tree. The default value of None would direct the algorithm to expand the tree until all leaves contain the least number of samples (the exhaustive approach) resulting in 99.75% accuracy for the training set and 75% for the test set. We then cycled through different depths from 1-9 and compared the accuracies. The model's accuracy increases as the depth of the tree increases but dips back down at 93.75% and 78.0% respectively at depth of 9. The chosen depth for the model was depth 5, due to the training and test accuracies being the closest at this point. This resulted in an 81% accuracy for the decision tree model.

Linear Regression

Linear regression is a statistical method that analyses and finds relationships between two or more variables. All the independent variables except for Salary were exploded variables from the categorical variables. The independent variables in the training and testing sets were scaled. The LinearRegression module of Scikit-learn was used to fit the training dataset and predict the

test dataset. We did this once with all features, with selected features, and with PCA. The PCA reduced the number of features to just one. Table 4 shows the test data and the predicted values in each case. The columns Test_cls, and PredPca_Cls are categorical representation of the test data and predicted values when using PCAfeature. We used these two columns to create a confusion matrix. The confusion matrix showed 65-68% range for classifying a correct category for predicting low, medium, and high versus actual low, medium, and high spenders. This model produced an overall accuracy of classifying 66.5% correctly.

	TestData	PredAllFeatures	PredPca	PredSelect	Test_cls	PredPca_cls
0	857	1295.534434	1061.925885	1493.792637	Med	Med
1	2191	1887.888907	1364.254187	1537.672161	High	Med
2	1071	1550.032086	1341.560353	1219.302853	Med	Med
3	983	1410.275587	1299.308579	1696.369370	Med	Med
4	1485	2265.459399	2286.929765	2319.889819	Med	High
...	...	...	...	...	...	...
195	1985	1821.640592	1940.478024	1554.137377	High	High
196	757	614.882612	1596.266268	733.132763	Med	Med
197	526	492.118473	1229.000783	671.297056	Med	Med
198	340	540.845802	1645.421998	688.728512	Low	Med
199	459	378.646566	193.035482	495.517380	Low	Low

Table 4: Linear regression predictions

Logistic Regression

Logistic regression is a model that can be used to predict the probability of a categorical dependent variable and is often used for binary classification. It builds linear models based on the log of the odds ratio which is the ratio of the outcome probabilities. Logistic regression requires that there are no missing values and that all inputs are numeric.

Scikit-learn's LogisticRegression was used to create this model using the csv file containing cleaned data. We observed an accuracy of 67.5% for the training set and 69% for the

test set. The model performed fairly well in predicting low and high spenders with correctly predicting over 70% for each. However, the model performed poorly for classifying medium spenders and only predicted about 31% correctly. It predicted over 59% of actual low spenders as medium spenders.

Random Forest

Due to the high accuracy for the decision tree model of 81%, a random forest model was also tested as our last model. A random forest is an ensemble model that is made up of many decision trees. Each decision tree works on its own to make a prediction of the class. The class that is most frequent is selected as the prediction for that sample.

To create this model, Scikit-learn's RandomForestClassifier was used. At first, a baseline random forest model was created with the default parameters that had an accuracy of 79.5%. Once a baseline model was established, we wanted to see if we could improve the model by changing the parameters. Scikit-learn's RandomizedSearchCV was utilized to cycle through different values for each parameter to find the optimal values. The values suggested by the RandomizedSearchCV resulted in 1,577 estimators, or individual decision trees in the ensemble model. These optimal parameters were used in the final random forest model.

The accuracy on the training set for the random forest model was 82%. This is a 1% increase from the decision tree model. The precision score returned the highest score for predicting high spenders. 83.67% of spenders were predicted high, when their assigned class was actually high. Predicting low spenders was also decent, with 85.59% predicted correctly. However, this score was lower than the decision tree model for low spenders. 70% was predicted correctly for medium spenders.

Results and Conclusion

Using different models, we can predict whether a customer is a low, medium, or high spender. In Table 5, we observe that the random forest and decision tree performed the best by far with an accuracy of 82% and 81% respectively. We can deploy this project in one of the company's direct mail marketing campaigns by targeting only high spenders identified by the random forest to reduce costs and increase profitability.

K-Nearest Neighbors	Decision Tree	Linear Regression	Logistic Regression	Random Forest
71.5%	81%	66.5%	69%	82%

Table 5: Model accuracies for predicting customer classes

We want to mention that there are some opportunities for improvement that should be considered before deploying this project. We recommend that the company collects additional customer information to improve model performance, specifically to increase the total number of rows in the dataset. In addition, the company may want to gather customer data on those who did not make purchases even after receiving catalogs. This will allow analysis to be done on take rate, which measures the percentage of customers acting upon receiving catalogs. By considering take rate, the company can allocate resources appropriately when trying to expand business to new buyers.

Acknowledgements

We would like to express our great appreciation for Dr. Brett Werner for his guidance and suggestions that were given throughout each milestone for this predictive analytics project. The advice helped us navigate the direction we wanted to go with the project. We would like to also extend our thanks to our classmates who contributed to our class discussions on the topics and techniques applied throughout this paper.

References

Abbott, Dean. (2014). *Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst.* Indianapolis, IN: John Wiley & Sons, Inc.

Analytics Vidhya. (n.d.). Getting Started With Decision Trees. Retrieved from https://courses.analyticsvidhya.com/courses/take/getting-started-with-decision-trees

Blackwell, J., DeCanio, T. (2009). Successfully Implementing Predictive Analytics in Direct Marketing. The Nature Conservancy, Arlington, VA. https://www.lexjansen.com/nesug/nesug09/sa/SA09.pdf

Breur, T., Paas, L. (2000). Building predictive models for direct mail: A framework for choosing training and test data. *J Database Mark Cust Strategy Management.* 8, 9–16. https://doi.org/10.1057/palgrave.jdm.3240013

Capobianco, T., Ibanez, C., Safari, E. (2020). *dsc630Project*. GitHub. Retrieved from https://github.com/cvibanez/dsc630Project

Galarnyk, Michael. (2017). Logistic Regression using Python (scikit-learn). Retrieved from https://towardsdatascience.com/logistic-regression-using-python-sklearn-numpy-mnist-handwriting-recognition-matplotlib-a6b31e2b166a

Koehrsen, Will. (2018). Hyperparameter Tuning the Random Forest in Python. Retrieved from https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74

Patil, Yogesh. (2018). Direct Marketing. Retrieved from https://www.kaggle.com/yoghurtpatil/direct-marketing

Yildirim, Soner. (2020). K-Nearest Neighbors (kNN) - Explained. Retrieved from https://towardsdatascience.com/k-nearest-neighbors-knn-explained-cbc31849a7e3