Direct Marketing

Preliminary Analysis

Milestone 3

Group 3

DSC 630

Torrey Capobianco

Conrad Ibanez

Edris Safari

Abstract

Predictive analytics is used to determine customer spending from receiving direct mailing catalogs. The company wishes to know which households are predicted to spend more from receiving a catalog, to increase their return on investment in direct mailing. This project analyzes the data the company has collected on their current customers. Predictive analytics is then used to present two models to aid in the business problem. We offer a model to predict how much a customer might spend and what type of spender they are predicted to be using machine learning regression and classification models respectively. Also explored are what characteristics contribute to a higher spending household through exploratory data analysis. The models presented to the company met our expected accuracy of 70% or higher. Put in place, we conclude that the models will be able to predict which households will be a higher spender so the company can direct their marketing material where interest is high.

Introduction

This project will utilize graphic and statistical learning techniques to perform predictive analytics on a direct marketing dataset. The goal is to understand the dataset, prepare it for modeling, evaluate and create appropriate models, test them, and measure their accuracy.

Background

Marketing is a business sector that can leverage predictive analytics to determine the type of customer to send their advertisements to. By seeing who will act on an advertisement geared towards them, companies can use customer data to focus their resources on targeting those that will take action vs. those that will not add significant sales from the marketing and distribution efforts. This project proposal looks at a company that conducts their business sales solely 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. On the other side of the spectrum, there are also those that will make significant purchases from receiving the mailer. Which households will act on these mailers?

Problem Statement

The company has collected data on previous customers that they have sent catalogs to. They wish to see how this information can help determine which households to send their mailers to in order to increase their return on investment. The problem this project addresses is which characteristics of a household will result in high sales. When given potential addresses, the company will be able to see who their higher spenders through our models would be. They can then focus their efforts on expanding their regions to customers that are likely to spend more than those that are not. 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, such as low, medium, or high spender. A recommendation of which model the company should use will be given. The remainder of the paper is divided into three sections, methods used for the predictive model, results from the models, and a closing discussion.

Data Understanding

Data Preparation

The data used in this project was retrieved from Kaggle. It contained 10 features and 1,000 samples. The amount spent feature is our target variable. This variable is used in different ways throughout this analysis to predict how much a customer might spend. 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. To prepare the data for this project, different cleaned data frames were created for the three different models: K-nearest neighbors, decision tree, and linear regression. Each model had a slight variation of input data. A category label was required for the classification models, which went through the feature creation process.

To prep the data for K-nearest neighbors (KNN), the amount spent was binned into three categories, low, medium, and high, which was below the 1st quantile, the inter quantile range, and above the 3rd quantile respectively. This became the classification labels for KNN. After dropping the history column due to missing values, all categorical variables were one hot encoded, exploding the variables to separate columns. Lastly, the numerical data was normalized, as it is the best method to find nearest neighbors with numbers on the same scale and not skewed.

To prep the data for linear regression, we took the similar approach as preparing data for KNN. We one hot coded the categorical variables but 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. We then moved the dependent variable 'AmountSpent' to the end of the table for better dissection of independent and depend variables in modeling section. The 'History' column was also dropped from the dataset due to missing data as well as our inability to explain why they were left blank. Those records with no history showed amount spent ranging from low to high, so there did not seem to be too much correlation. The resulting dataset with the following records was written to a csv file for the consumption of the linear regression modeling program.

```
Index(['Salary', 'Children', 'Catalogs', 'Old', 'Young', 'Gender',
'OwnHome', 'Married', 'Distance', 'AmountSpent'], dtype='object')
```

For data preparation for the Decision tree modeling, we took the same approach as KNN with the exception that we chose the following ranges for low, med, high categories:

Low: greater than MinQ, less or equal to MedQ

Med: Greater than MedQ, less than or equal to ThirdQ

High: Greater than ThirdQ, less than or eqaul to MaxQ

The resulting dataset with the following records was written to a csv file for the consumption of the Decision tree modeling program.

```
Index(['Salary', 'Children', 'Catalogs', 'Middle', 'Old', 'Young', 'Female',
'Male', 'Own', 'Rent', 'Married', 'Single', 'Close', 'Far',
'Amt_Spnt_Class'], dtype='object')
```

Exploratory Data Analysis

The amount spent is the independent variable for this project. The other nine variables will be evaluated in our predictive models. The dataset did not have much documentation, so we performed exploratory data analysis to get a better understanding and to get further insights. 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 are 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.

We created various graphs as part of our EDA including a correlation heatmap, histograms, boxplots, and scatterplots with regression lines. In Figure 1, the correlation heatmap indicates a strong positive relationship between the amount spent and salary. This is also confirmed in the scatterplot with a regression line in Figure 2. The heatmap also shows a positive relationship between amount spent and number of catalogs, which is also reflected in the scatterplot in Figure 3.

Figure 1 – Correlation Heatmap

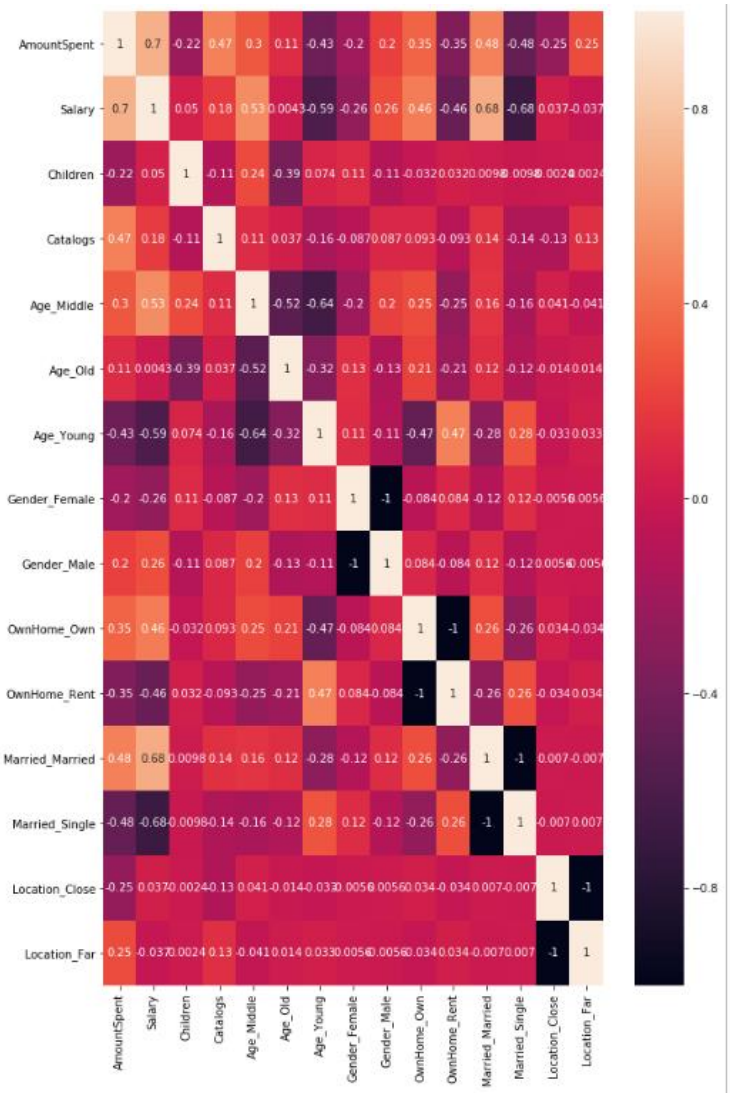


Figure 2 – Amount Spent vs. Salary with Regression Line

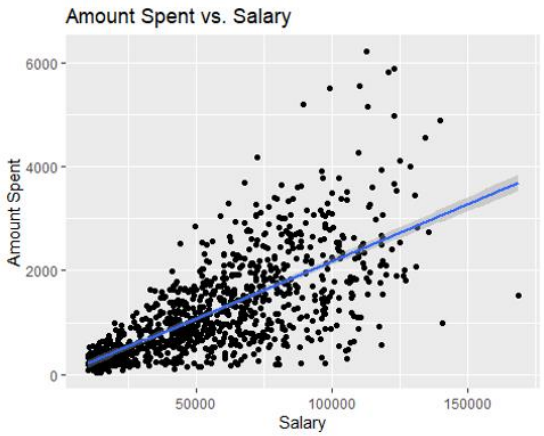
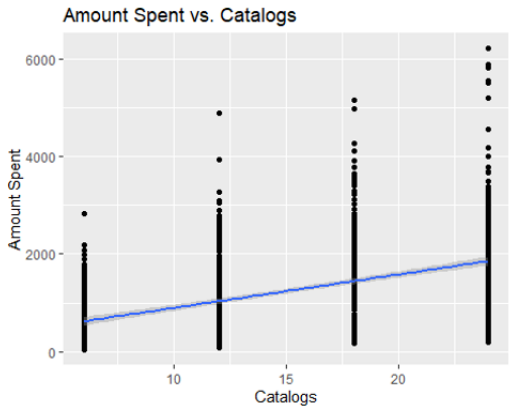


Figure 3 – Amount Spent vs. Catalogs with Regression Line



The heatmap also shows a strong correlation between those who are married and salary. This is also indicated in the scatterplot for amount spent vs. salary along with the married status

as labels in Figure 4. 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 shown in Figure 5. Our exploratory data analysis indicates people who are married and middle-aged, own a home, and have less children may spend more.

Figure 4 – Amount Spent vs. Salary with Married Status

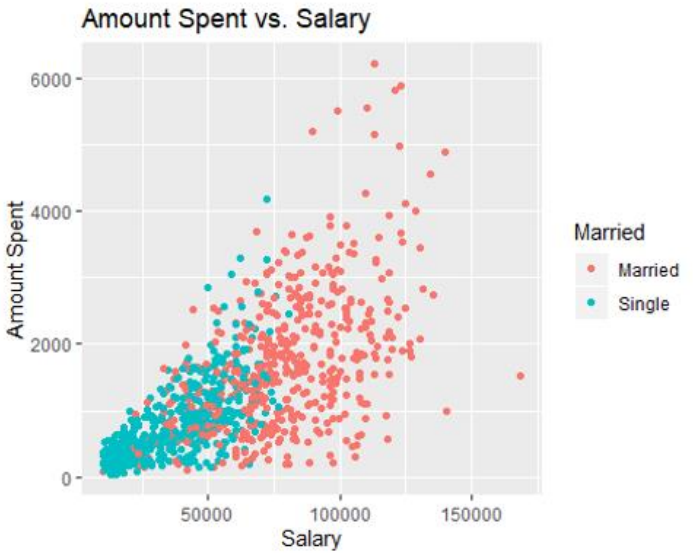
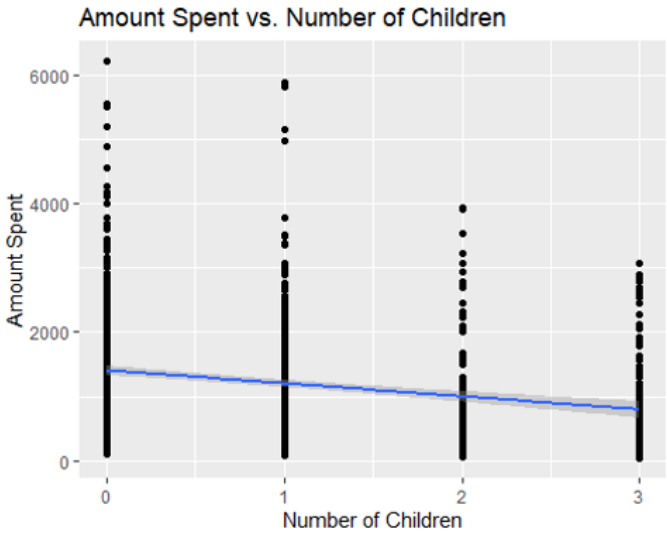


Figure 5 – Amount Spent vs. Number of Children



Methods

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. Through the data preparation stage, we grouped the samples into three types of spenders, low, medium, and high. The KNN classifier model is designed to predict which group the person belongs to by looking at the person's nearest neighbor's class. If its neighbors have similar data points to a person, they are most likely

to have similar spending behavior. The optimal values K, the number of nearest neighbors to pick, was decided from Scikit-learn's GridSearchCV function. This function tested 20 nearest neighbors cycling through 5 cross-validation data splits. The optimal neighbors for our model was 6 nearest neighbors.

The KNN model was first created with all the features except for history, as this was dropped in the data preparation phase. Using a train and testing set, with 25% of the data reserved for testing, the six nearest neighbors model resulted in 69.6% accuracy. This model was tested again but with cross-validation of 5 times and had a 73.6% accuracy. These results were lower than we expected, and a large varying degree in accuracy based on the different test set methods. A second model was created to remove number of catalogs, due to the uncertainty that our team had in the meaning behind this variable and how would it affect new customers, who had never received a catalog before. With this variable removed, the optimal number of k-neighbors was 4. The train/test model produced 68.8% accuracy and the cross-validation model, 73.1% accuracy. The accuracy was roughly the same, with a slight decrease with this KNN 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. While a simple linear regression between salary and amount spent would make sense, we decided to take the all-in approach as a 1st pass. We divided the dataset to 80% for training 20% for test data sets. We then scaled the independent variables in both datasets. We then fitted the training data set using the LinearRegression module of Sklearn., the

predicted values were then calculated using the test data set. Comparison of test data and predicted shows a wide variance between the test data used and the predicted values.

:

	TestData	PredData	difference	Difference
0	857	1295.534434	438.534434	438.534434
1	2191	1887.888907	-303.111093	-303.111093
2	1071	1550.032086	479.032086	479.032086
3	983	1410.275587	427.275587	427.275587
4	1485	2265.459399	780.459399	780.459399
5	1634	1510.618763	-123.381237	-123.381237
6	917	1348.905668	431.905668	431.905668
7	1120	1056.215892	-63.784108	-63.784108
8	821	1112.635339	291.635339	291.635339
9	2574	2025.540675	-548.459325	-548.459325

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 nod. Then the algorithm splits the node to child nodes and the child nodes to their respective leaf nodes are reached. The measure and strength of the splits are evaluated and optimized by the decision tree algorithm. We chose the Gini algorithm for this dataset because the target variable is categorical, and the predictors are mixed.

Sklearn's DecisionTreeClassifier, can be used to train and test the dataset. It's 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). Running the model with this setting resulted in the training accuracy of 99.73% and test accuracy of 73.2% for the test dataset. We then ran the model with depth of 1-10 and

compared the training and test accuracy. The result shows that the accuracy increases with the depth of the tree and max out at 99.73% and 73.2% respectively.

	max_depth	train_acc	valid_acc
0	1	0.620000	0.600
1	2	0.660000	0.696
2	3	0.710667	0.704
3	4	0.765333	0.740
4	5	0.801333	0.752
5	6	0.844000	0.748
6	7	0.876000	0.756
7	8	0.902667	0.756
8	9	0.925333	0.764

Results

During the course of this study, we used python and R to perform EDA and data preparation and ran the dataset through three different models. The statistical results of each would indicate that we must pursue further into feature selection and determine which set of features produce best results. The dataset itself could be augmented with additional and more pertinent predictors such as education, job type, interests, and mostly the type of merchandize they purchased.

We believe that the study itself is warranted but can be affective if the dataset were to expand as suggested.

Discussion

The dataset for this project is not very large and does not have a lot of features. There is not much collaborative work for it in Kaggle, and we are not sure whether it is real-world data. It also does not have a lot of documentation nor does it have many details about the data. For example, there is no information on the type of store or products that are sold to the customers. For certain attributes, we are making some assumptions. Catalogs appear to be in multiples of

six, so perhaps six catalogs are mailed out at a time. Additionally, we are not quite sure on the exact meaning for the location and history attributes.

Only details for customers who made purchases are provided. No details on the total amount of direct mailings or customers who received mail and did not make any purchases are given. Therefore, we are unable to determine and predict metrics such as take rate. Our primary focus for this project has been to apply different predictive models for the amount spent and evaluate their performance. We will look at additional ways to apply predictive analytics.

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

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

GitHub Repository: https://github.com/cvibanez/dsc630Project