Report/White Paper

Direct Marketing

DSC 680

Saurabh Shrestha

Business Problem

The problem this project addresses is which characteristics of a household will result in high sales. 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. When given potential addresses, the company will be able to see who their higher spenders through our models would be. Through predictive analytics, I present multiple models to identify how much a person is predicted to spend and what type of spender I would label them as, such as low, medium, or high spender.

Background/History

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?

Data Explanation

The data used in this project was retrieved from Kaggle. The data set includes n = 1000 customers and the following variables: Age (of customer; old/middle/young); Gender

(male/female); Own Home (whether customer owns home; yes/no); Married (single/married); Location (far/close; in terms of distance to the nearest brick and mortar store that sells similar products); Salary (yearly salary of customer; in dollars); Children (number of children; 0–3); History (of previous purchase volume; low/medium/high/NA; NA means that this customer has not yet purchased); Catalogs (number of catalogs sent); and Amount Spent (in dollars).

Methods

Handling missing data is important as many machine learning algorithms do not support data with missing values. First I need to work on if there are duplicates or missing values. For data cleaning and pre-processing I would check and deal with missing and duplicate variables from the data set as these can affect the performance of different machine learning algorithm. Additionally, I would work on creating boxplot and treat outliers. Such as I will remove the outliers as anything over 75% of the maximum value. Additionally, perform Exploratory Data Analytics - Here I would to gain important statistical insights from the data such as distributions of the different attributes, correlations of the attributes with each other and the target variable.

Modeling

For the machine learning algorithm, I would scale the features to speed up the training of the classifiers and then split the data into a training and test set at a ratio of 0.8 to 0.2 respectively.

Several algorithms could be used for machine learning model. For this project I would chose following models: Logistic regression because it models the probability of a data point belonging to a particular class and assigns this point the appropriate label based on a chosen threshold. K-Nearest Neighbor because the algorithm assumes that similar data points that are near each other will belong to the same class. Linear Regression because is a statistical method that analyses and finds relationships between two or more variables. It would be helpful to work with it on salary and amount spent variables.

After training each model and tuning their hyper-parameters using grid search, I would evaluate and compare their performance using the following metrics: **The accuracy score:** which is the ratio of the number of correct predictions to the total number of input samples, The F1

Score: Which is defined as the weighted harmonic mean of the test's precision and recall, The Area under the ROC Curve (AUC): Which provides an aggregate measure of performance across all possible classification thresholds.

Analysis (with Illustrations)

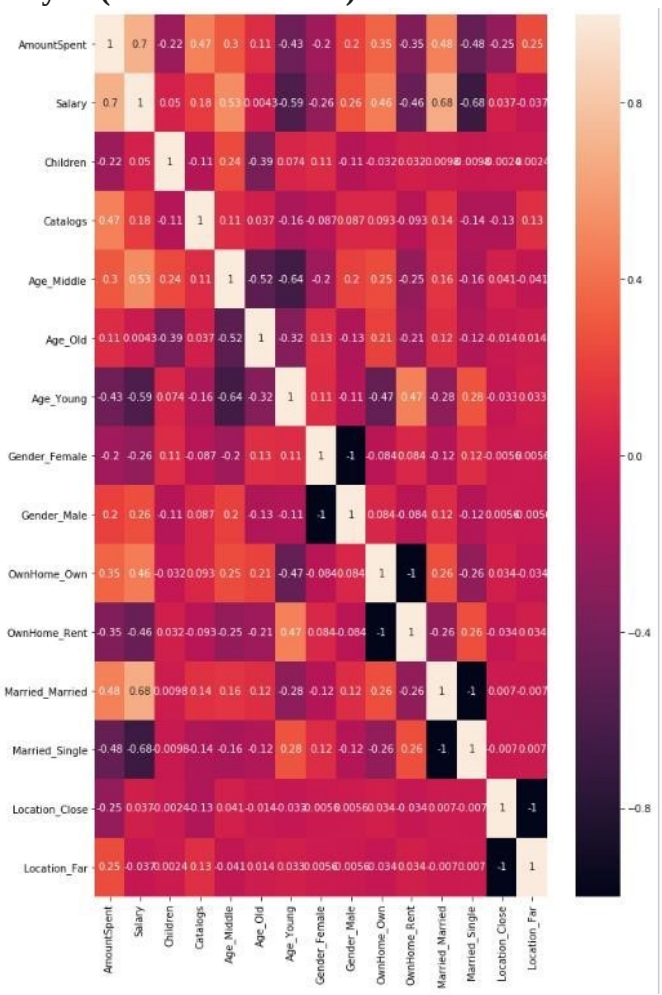


Figure 2– Amount Spent vs. Salary with Regression Line

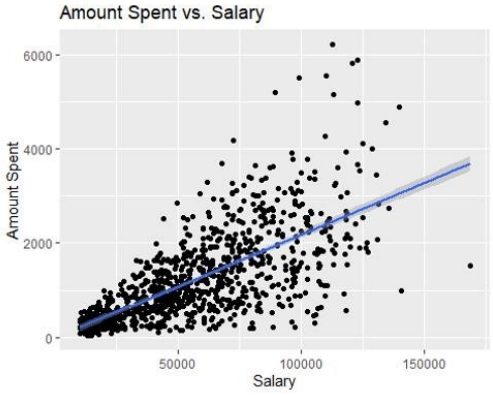
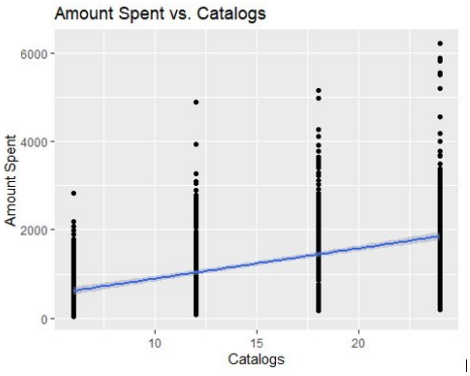


Figure 3– Amount Spent vs. Catalogs with Regression Line



Observations

- **In Figure 1**, the correlation heat map 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 heat map also shows a positive relationship between amount spent and number of catalogs, which is also reflected in the scatterplot in Figure 3.

- **Additionally**, the heat map shows a strong correlation between those who with age and salary. This is also indicated in the scatterplot for amount spent vs. salary along labels in Figure 4.

Conclusion

During the course of this study, I used python to perform EDA and data preparation and ran the dataset through three different models. The statistical results of each would indicate that I 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.

Assumptions and Limitations

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 I am 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, I am making some assumptions. Catalogs appear to be in multiples of six, so perhaps six catalogs are mailed out at a time. Additionally, I was not quite sure on the exact meaning for the location and history attributes. Limitations include 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, I was 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.

Challenges

Issues which could be faced while working on this project would be:

1. Data size too small to apply prediction algorithm.
2. Outliers/Skewness in data.
3. Prediction model not returning expected results.
4. Learning curve.
5. Not enough impactful variables present in dataset.

Recommendations

I would suggest this project would be helpful primarily to marketing companies and retailers who specializes in direct sales. With the help of analysis, companies can dissect demography, spending age brackets, understand behavioral patterns and to learn what makes some customers spend more than others.

Implementation Plan

The implementation plan contains the below steps to accomplish the project.

- Business Understanding
- Data Preparation
- Exploratory Data Analysis
- Data preparation for modelling
- Modelling
- Model Evaluation

Ethical Assessment

The reason for ethical assessment is to figure out what to do next. Ethics means the consideration of behaviors that will do good and assessment is what you do to figure out what you or someone else knows about doing a particular thing—and then figuring out what the next step is. For this project, I researched when working with the variables, followed the rules and regulations of ethics such as how could my system negatively affect impact customers? who is the most vulnerable and why?