Direct Marketing

Milestone 2

DSC 630

Torrey Capobianco

Conrad Ibanez

Edris Safari

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. As a final step, all findings will be consolidated in a presentation and a document.

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 through the use of 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. Through predictive analytics, we

will create a model that will identify the attributes that will predict higher spending from a customer. The model created will identify those households that the company should direct their mailers to. They can then focus their efforts on expanding their regions to customers that are likely to spend more than those that are not.

Scope

To achieve this, the scope of this project requires several steps. The data will go through a preparation phase, exploratory phase, model creation, and validation. We will require three weeks to complete this to have a preliminary analysis deliverable at the end of that timeline. Seven weeks out, intermediate results of the model will be delivered with a final presentation of our findings in ten weeks. The remaining project proposal goes over the data source, planned approach, model testing and deployment, expected results, and the execution of this project.

Preliminary Requirement

As a preliminary requirement, we will secure a dataset with ample features and a good set of candidates for target variables. We have chosen the Direct Marketing dataset in Kaggle. We will decide on the programming languages to use as well as which to use for what purpose (i.e. R for graphics, data understanding, and preparation and possibly switch to python for modeling, test, and validation.

Technical Approach

Analysis

Categorical variables will be encoded, and the selected features will be scaled accordingly to create a uniform data set. Features will then be analyzed by performing summary statistics and creating graphs such as bar, scatter and histograms. Scatter plots and correlation matrices will help us determine if we can see any pre-existing correlations that we might

determine that is a predictive variable in the model. Histograms can help us see the frequency of specific characteristics of the customer base. This will all help us perform an appropriate analysis to find, create, and assemble features that can best be fitted into a model.

Requirement Development

The focus of this exercise is to predict amount spent and to predict a binary yes will purchase and no will not purchase. The features such as gender, distance, and income will be used in the models to see if an accurate prediction can be made.

Model Deployment

We will use unsupervised learning methods to determine relationships between different variables. The following descriptive modeling algorithms will be considered for this project:

- Principle Component Analysis (PCA)
- Clustering algorithms.

We will apply supervised learning methods to determine relationships between inputs to the two target variables, whether the customer purchased anything and if so, how much did they purchase. The following predictive modeling algorithms will be evaluated:

- Decision Trees
- Logistic Regression
- K-Nearest Neighbor
- Naïve Bayes
- Linear Regression

Testing and Evaluation

There are various ways to test the accuracy of models. We will look into the traditional way of splitting the dataset into 30% for testing and 70% for training the dataset. However, due

to number of samples in the dataset being 1,000 records, we may also look into testing the Bootstrap Sampling method to train and test the models. To assess the predictive models, we will calculate the percent correct classification (PCC) and create confusion matrixes. The metrics of accuracy, recall, and precision will be calculated from the confusion matrix.

Expected Results

We hope to make predictions at greater than 70% accuracy. However, with lack of untested and untraining data, the accuracy cannot be ascertained.

Management Approach

Project Plan

We will complete project milestones and team tasks per the below plan:

Week 1: Milestone 1 Due (Team Information/Communication Plan) - Completed

Week 2: Milestone 2 Due (Data Selection and Project Proposal) – Current Week

Weeks 3-4: Data Preparation, Exploratory Analysis (graphs, statistics), Descriptive Modeling, and Initial Predictive Modeling and Testing

Week 5: Milestone 3 Due (Preliminary Analysis)

- Deliverable: A report for several models that have been created and tested

Weeks 6-8: Continue with Predictive Modeling and Testing, and Evaluation of Predictive Models

Week 9: Milestone 4 Due (Project Presentation & Status)

- Deliverable: Presentation deck geared towards the company's executives

Weeks 10-11: Complete all coding, Code Review, Finalize Paper

Week 12: Milestone 5 Due (Final project paper and presentation)

- Deliverable: Finalized predictive model with accuracy score and conclusions presented in a paper and presentation for the company

Project Risk

The dataset does not have much prior work, analysis, or collaborators. Therefore, it is unclear if predictive modeling will work well on this dataset and if any valuable insights will be gained by this project. The dataset also has too many categorical variables. We may need to augment with some quantitative variables such as liabilities (loans, mortgages, etc.). Another project risk is that the dataset is rather small with 1000 records; however, this size should suffice our intention of applying the techniques.

We will be completing this project as a team of three individuals located in three different time zones, so meeting and collaborating will be a challenge. We are mitigating this risk by being flexible and meeting at least once a week. Project work will be divided equally among the team to promote independent work.

References

1. Abbott, D. (2014). Applied Predictive Analytics. Indianapolis, IN: John Wiley & Sons, Inc.
2. Direct Marketing dataset. https://www.kaggle.com/yoghurtpatil/direct-marketing