ISYE 7406 Project Proposal

Online Shopping Purchase Intent

March 14, 2022

1 Your Names

Casey Stanfield, Jia Hui Mo, Mariah Carter, Simon Yee, Xu Anne Zhang

2 Project description

Our goal is to predict whether or not revenue was made based on various purchase intent factors including the number of pages visited per category, visitor type, and other e-commerce website Google analytics metrics.

The categories of web pages include informative, administrative, or product related. Other predictor variables include information about the region, browser type, whether the visit was during a weekend, if the visit is close to a holiday, and the month of the visit. Specific google analytics metrics include bounce rate and exit rate.

The focus of this project is to infer the relationship between whether revenue was made and these predictor variables using various data mining and statistical learning methods. We will use several different binary classification methods to predict revenue and compare these different methods with one another.

3 The Dataset

For this project our team has decided to use the "Online Shoppers Purchasing Intention" Dataset sourced from the UCI Machine Learning Repository. We obtained this data online through their website.

- The dataset is publicly available at the following Link (Click Here).
- The url: http://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset

4 Proposal

4.1 Scientific Research Questions To Address

There are several past articles and academic papers in relation to this dataset that showcase the use of advanced Neural Networks (NN), ensemble methods such as Random Forest (RF), and dimensionality reduction techniques such as Principal Component Analysis (PCA). To differentiate this project analysis our focus is on applying different modeling and data mining techniques to attempt to provide an alternative to Random Forest with an approach that achieves a similar accuracy level and more interpretability than possible with NN or PCA.

We are interested in finding justification and inference on each method and comparing it with common research methodologies such as Random Forest. Are the uses of certain models like Random Forest across multiple research papers reasonable with our data set? Do their results make sense? How

do other models respond? How do the variables interact quantitatively and qualitatively? To address these questions, we will conduct new data mining and statistical learning methods.

4.2 Data Mining Statistical Learning Methods

A common research methodology found to be used with this dataset in past research is Random Forest, and as an example, we would like to see if we can justify its use, that is, are the relationships between variables so complicated that we cannot get a "reasonable" prediction with better interpretability. We plan to use methods such as discriminant analysis and smoothing to understand if we're able to build simpler and more efficient models that can help us better understand the relationship between the predictor variables and revenue. We may also choose other classification and probability models that are not commonly chosen in other papers to see if we can infer what or where the challenges are.

As stated above this project aims to differentiate itself by exploring different modeling and mining techniques to achieve a more interpretable model with high accuracy. Our project will explore data mining techniques such as feature engineering and feature selection to attempt to attain greater predicting power while removing variables with little impact to the model accuracy. On the modeling side, to avoid the less interpretable models, we will explore models including Logistic Regression, Naïve Bayes, LASSO, Discriminant Analysis, Kernel Smoothing, and various ensemble methods. We hope that these methods will help us to frame the analysis around business use cases such as helping to guide decisions around how to best structure e-commerce websites to funnel traffic towards sales or when to run promotions to drive the greatest return on marketing dollars spent during certain times such holidays.

References

Kabir, M. R., Ashraf, F. B., amp; Ajwad, R. (2019). Analysis of different predicting model for online shoppers' purchase intention from Empirical Data. 2019 22nd International Conference on Computer and Information Technology (ICCIT). https://doi.org/10.1109/iccit48885.2019.9038521

Vigneshprakash. (2020, October 1). Online shoppers purchasing intention (PCA,smote). Kaggle. Retrieved March 13, 2022, from https://www.kaggle.com/vigneshprakash/online-shoppers-purchasing-intention-pca-smote

Dissanayake, I. (2019, October 7). Online shoppers purchasing intention: Randomforest: ML. Medium. Retrieved March 13, 2022, from https://medium.com/analytics-vidhya/ospi-mul-randomforests-156acdb73fd9

Dharmasiri, M. A. (2019, October 8). Preprocessing data for predicting online shoppers purchasing intention: ML. Medium. Retrieved March 13, 2022, from https://medium.com/analytics-vidhya/preprocessing-data-for-predicting-online-shoppers-purchasing-intention-ml-ba78186b7e85