Social Network Ads Analysis

Sumit Patel

Springboard Capstone #1

Abstract

Social network ads are commonly deployed due to their cost effective nature and ease of deployment. In the observed dataset we will observe the use of specific parameters and their predictive probability on purchases. In order to have a successful campaign the parameters need to correlate to purchases as well as be statistically significant. It is important that the ads be deployed more specifically to users that would be more likely to make a purchase than not. We will be analyzing a Kaggle dataset consisting of 400 rows and 5 columns. The columns pertain to User ID, Gender, Age, Estimated Salary and Purchased. Through exploratory data analysis we will identify the optimum parameters that impact purchase outcomes. The findings from EDA will then be used to build models to see the most optimum fit for ideal results.

Kaggle Dataset can be found at:

https://www.kaggle.com/rakeshrau/social-network-ads/data

Social Network Ads Analysis

Social network ads are commonly deployed due to their cost effective nature and ease of deployment. In the observed dataset we will observe the use of specific parameters and their predictive probability on purchases. In order to have a successful campaign the parameters need to correlate to purchases as well as be statistically significant. It is important that the ads be deployed more specifically to users that would be more likely to make a purchase than not.

Method

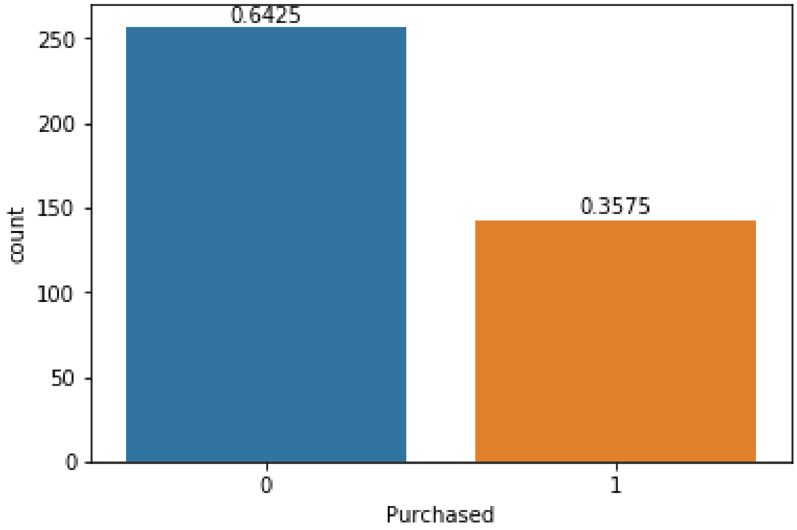
Data

Dataset is a Kaggle dataset that consists of 400 rows of unique customer data and 5 columns consisting of attributes: User ID, Gender, Age, Estimated Salary, Purchased. The dataset can be found at: https://www.kaggle.com/rakeshrau/social-network-ads/data. Our observations will provide that the dependent variable be Purchased as that is the value that we want to observe and optimize for the campaign.

Exploratory Data Analysis (EDA)

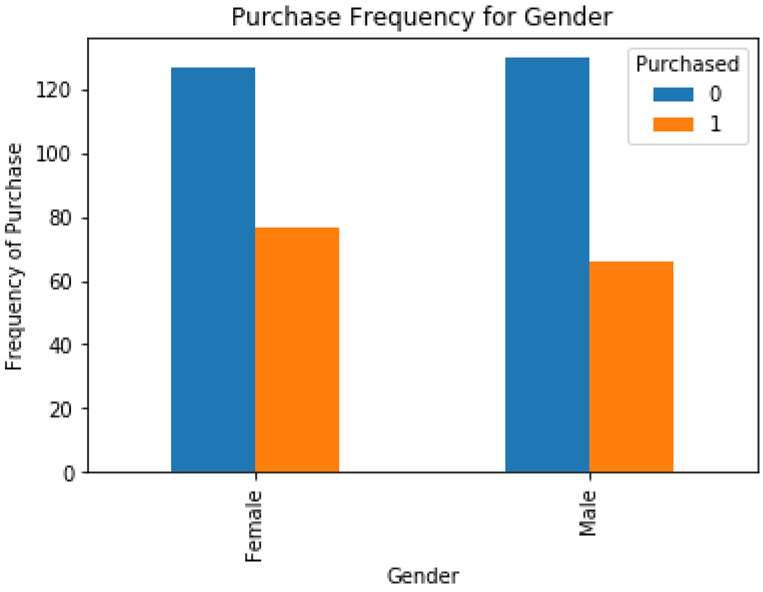
Our initial observations on the data to find values that correlate to the Purchased outcome with the best measure. Initial EDA observations against the variables of Gender, Age and Estimated Salary to find the best correlation to Purchased was needed to create models with the best fit for the campaign. In order to understand the data and any patterns that may be presented. Initial observations were to investigate how purchase patterns and behaviors were with the collected sample.

Figure #1 Purchases count with percentage



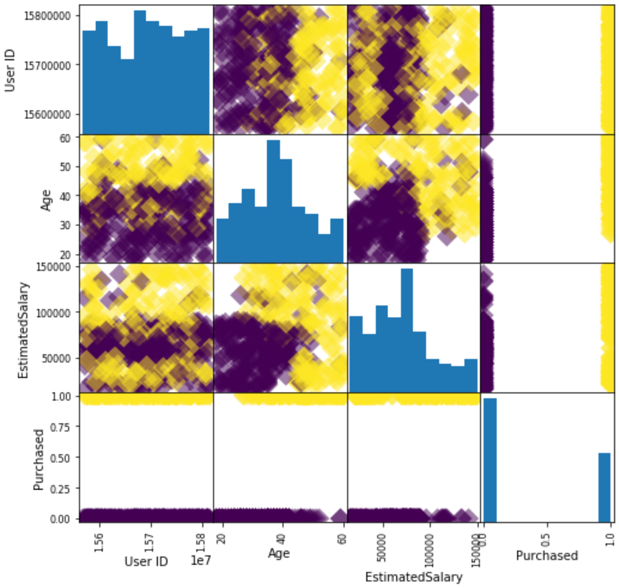
From the above figure we can see the current ratio of purchases versus non-purchases. Purchases are significantly less than non-purchases.

Figure #3 Purchase Frequency for Gender



We can see that Gender does not significantly impact Purchased for our observations and can be excluded from our models.

Figure #3 Correlations of all variables to Purchase



From this figure we can see the correlations of the variables against Purchased. We can also identify the categorical variables bi-sectional groupings. We also see strong correlations for Age and Estimated Salary.

Model Selection

After extracting the necessary data and processing the attributes necessary for our models we create a training set and testing set to make our observations with our models. We will be using Logistic Regression, Elastic Net, K-Nearest Neighbors and Support Vector Classifier as methods for our models. Observations and findings will be reported in results. We will be observing the scores on the testing sets for our observations. From this we can see that the highest accuracy score is for KNN with Standard Scaler yielded the highest accuracy score of 94%. Overall value the best model was SVC with Standard Scaler with accuracy score of 90% and lowest MSE value. This model also had the highest precision, recall and f1 scores. The worst model was the logistic regression.

Results

The results of the models are as follows:

Classifiers	Precision	Recall	F1-score	Accuracy score	Mean-squared Error
Logistic Regression	**0.00**	**0.00**	**0.00**	**0.70**	**0.55**
Logistic Regression with GridSearchCV	**0.00**	**0.00**	**0.00**	**0.84**	**0.55**
K-Nearest Neighbors (n=5)	**0.82**	**0.79**	**0.81**	**0.86**	**0.37**
K-Nearest Neighbors (n=3)	**0.85**	**0.76**	**0.80**	**0.86**	**0.37**
KNN with Standard Scaler	**0.38**	**0.38**	**0.38**	**0.94**	**0.61**
Support Vector Classifier	**0.67**	**0.17**	**0.27**	**0.73**	**0.52**
SVC with Standard Scaler	**0.86**	**0.79**	**0.83**	**0.9**	**0.32**

For the purpose of this study testing scores were observed from classification report. Accuracy, precision, recall, f-1 score and mean squared error were all used for determination of best model. Per the objective of the study, threshold probability for classification of Purchased needed optimization to increase probability for purchase outcome from ad deployment. Under this condition Logistic Regression, KNN and SVC were used accordingly. Per the findings SVC with Standard Scaler proved to be the best model for our purposes. SVC when used in conjunction with Standard Scaler had the smallest MSE and best scores overall making it the ideal model for purchase outcome relation to our social network ad deployment success.

References

Raushan, Rakesh. *Social Network Ads* | *Kaggle*, 6 Aug. 2017, www.kaggle.com/rakeshrau/social-network-ads/data.