# 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 of features. The features are 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

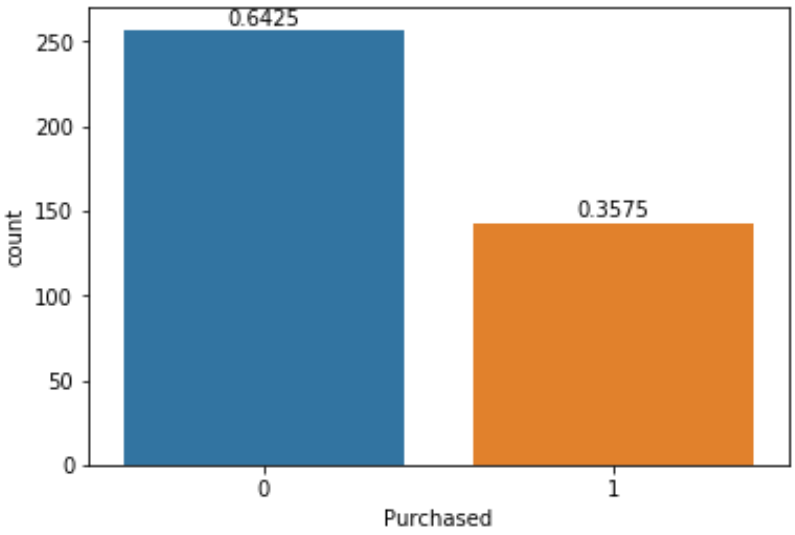
### Data

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)

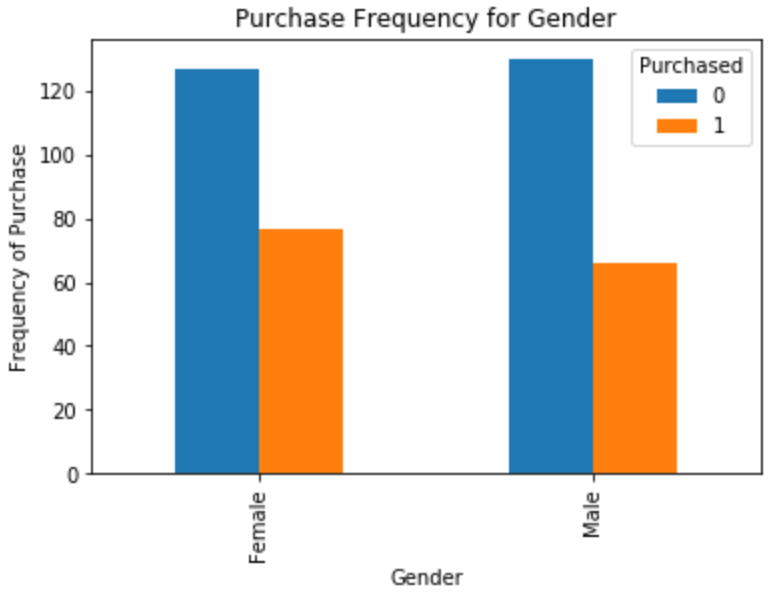
Our initial observations on the data to find values that correlate to the Purchased outcome with the best measure. Initial EDA observations against the variables of Gender, Age and Estimated Salary to find the best correlation to Purchased was needed to create models with the best fit for the campaign. 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



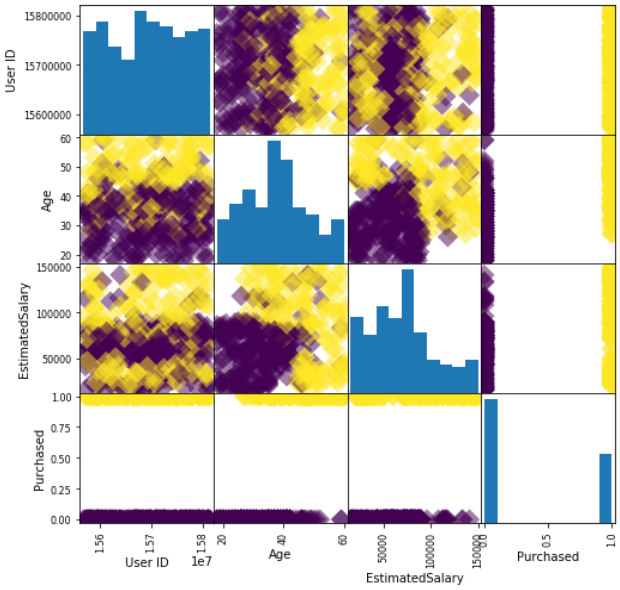

*From the above figure we can see the current ratio of purchases versus non-purchases. Purchases are significantly less than non-purchases.*

**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.*

## Feature Selection

We choose Gender and Estimated Salary as our features of interest due to their correlative significance. We can see from Figure 3 that Age and estimated Salary both provide a distinct segmentation on the correlation where we see two very distinct groupings. This is also apparent when compared with User ID but becomes ambiguous and not as distinct. Age and Estimated Salary are good features to use for deployment as the characteristics are viable for the ad deployment. As we have seen from the EDA, users of certain age groups and estimated salaries are likely to make purchases.

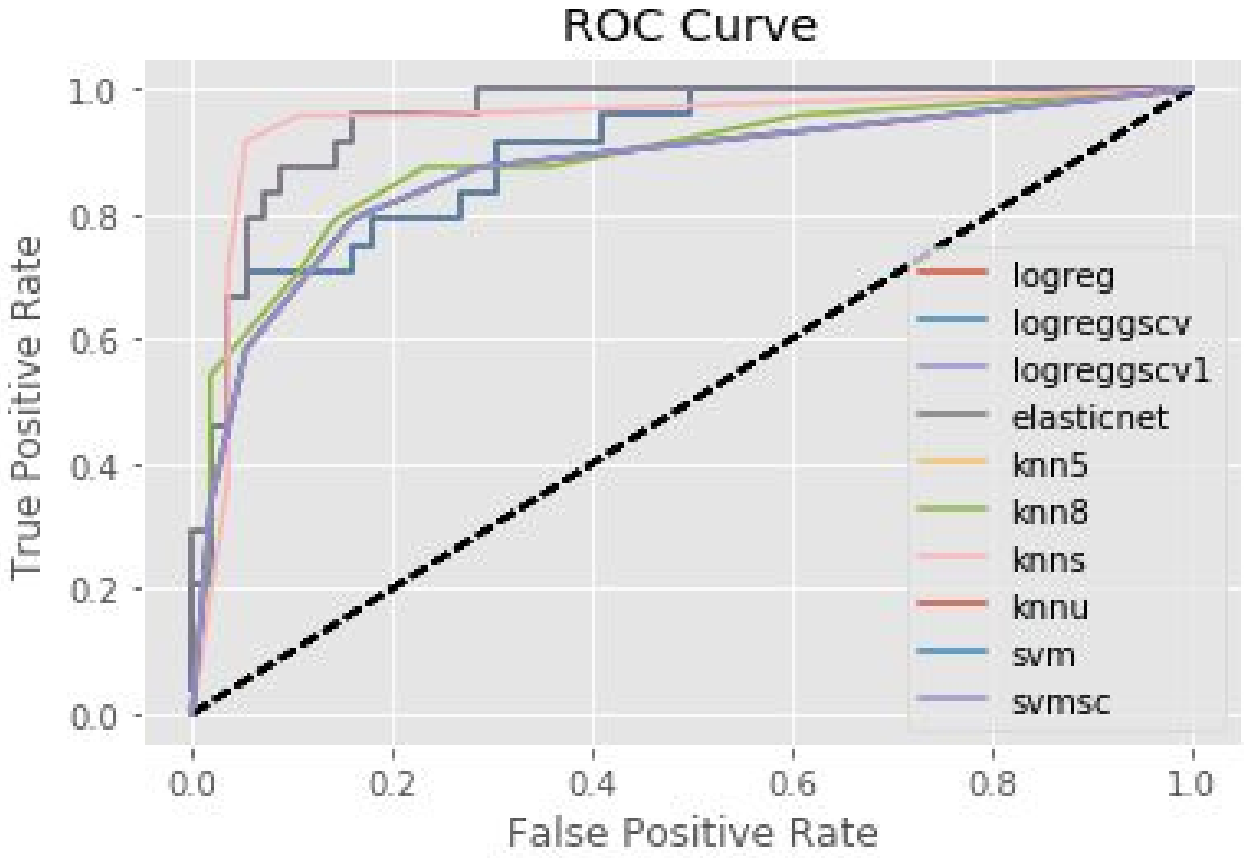
**Model Selection**

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. Metrics that will be observed are the classification reports as well as the AUC scores of the ROC curves. The ROC curves provide a good visual performance metric as to the performance of the model and it's predictive capabilities. The AUC provides the value for the ROC and range between 0.867 and 0.952 which relates that our models have fairly good predictive measures overall. The deterministic feature would be the goals of the client and how they would like to use the predictive metrics provided.

**Results**

The findings of the models are as follows:

| Classifiers | Precision | Recall | F1-score | Accuracy score | AUC |
|---|---|---|---|---|---|
| Logistic Regression | 0.09 | 0.30 | 0.14 | 0.30 | 0.8973 |
| Logistic Regression with GridSearchCV | 0.09 | 0.30 | 0.14 | 0.70 | 0.8973 |
| Logreg with GSCV and params | 0.09 | 0.30 | 0.14 | 0.84 | 0.952 |
| Elastic Net | 0.09 | 0.30 | 0.14 | 0.84 | 0.952 |
| K-Nearest Neighbors (n=5) | 0.86 | 0.86 | 0.86 | 0.86 | 0.905 |
| K-Nearest Neighbors (n=8) | 0.84 | 0.84 | 0.83 | 0.84 | 0.906 |
| KNN with Standard Scaler | 0.62 | 0.62 | 0.62 | 0.94 | 0.943 |
| Support Vector Classifier | 0.71 | 0.72 | 0.66 | 0.73 | 0.867 |
| SVC with Standard Scaler | 0.90 | 0.90 | 0.90 | 0.90 | 0.867 |


ROC Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
logreg
logreggscv
logreggscv1
elasticnet
knn5
knn8
knns
knnu
svm
svmsc

**Conclusion**

Evaluations of the models primarily used the classification report scores and ROC in conjunction with AUC score. The classification report provides insights to performance of model for ability to correctly identify labels. ROC in conjunction with AUC is used to determine the performance of the models and how random or consistent the predictions are. Our results determined that logistic regression with gridsearchCV using parameters or Elastic Net produced the most ideal results for predictive analysis on social network ad deployment success. Logistic regression is the best model as long as it is tuned with either gridsearchCV and extended parameters or ElasticNet as they both yielded the same result. They both share the same exact performance level in the ROC curve figure. The logistic regression has the better AUC score relating that the true positive rate in this model is higher that the other models relating that the predictive significance to the dataset is better pronounced in this model with less errors. The ROC Curves figure shows the performance of each model and the visual representation of the AUC score. This relates that on a random consumer the outcome prediction being a true positive result is within 95% accuracy. Relating to the scenario this means that for a random user that the ad is deployed to the certainty that they will produce a true positive result, in this instance a purchase, is within 95% likelihood.

# References

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