Capstone Project Presentation

To Predict Purchase Outcome of Social Ad on Consumer

Sumit Patel

Objective of Social Network Ad

- The objective of a Social Network Ad is to take certain parameters of interest to optimize to consumers for a better likelihood outcome which in this context would be purchases.

- The initial step is to identify the variables of interest that relate most accurately to purchase outcomes. Using the dataset with observed instances of users that Purchased with outcome of ‘0’ for non-purchases and ‘1’ for purchased.

Prediction Outcomes for each Observation:

- **Case 1: True Positive**
    - **Predicted as ‘1’ and is actually ‘1’**
- **Case 2: False Positive**
    - **Predicted as ‘1’ and is actually ‘1’**
- **Case 3: True Negative**
    - **Predicted as ‘0’ and is actually ‘0’**
- **Case 4: False Negative**
    - **Predicted as ‘0’ and is actually ‘1’**

- The cost associated with the deployment of the social network ads campaign would be C dollars (denoted for cost). Revenue earned for each ad would be R dollars (denoted for revenue). Assuming R = 2C and payoff for each case would be:
  - **Case 1:** R - C
  - **Case 2:** **- C**
  - **Case 3:** 0
  - **Case 4:** - (R - C)
- The campaign needs a model that can identify **Case 1** and **Case 4** with greater accuracy while reducing the instances of **Case 2** and **Case 3**.

Purpose of the Model

- Model selection needs to be done to find a model that is most suitable to the data. Steps to be considered:
  - Pre-processing of data
  - Create a training and test set from data. Size of sample out of dataset used will be 20%. I am using both 3 and 5 fold cross validation.
  - Create a pipeline
  - Fit and predict on different classes of models to observe highest accuracy for our requirements.
  - Measure different evaluation metrics for the identification of the most suited model(s).
  - Find models that can reduce the probability of instances for **Case 2** and **Case 3** to reduce cost and maximize profitability.

Data Description

- This dataset originates from a Kaggle dataset comprised of 400 rows with attributes of User ID, Gender, Age, Estimated Salary and Purchased.
  - Gender is a categorical variable comprised of ‘Male’ and ‘Female’ attributes
  - Purchased is also a categorical variable comprised of ‘0’ indicating non-purchase and ‘1’ indicating purchases
- The binary classification goal is to predict the outcome probability of the user to make a purchase.
- After preprocessing the total observations used were also 400 with nothing lost in the preprocessing. The data is weighted more towards non-purchases as opposed to our goal of purchases. Only 35.75% of observations have made purchases.

Preprocessing and Model Evaluation

The following was performed:

- Creating dummy variable for categorical values of Gender variable
- Identifying the relevant independent and dependent variables based on their correlation with the dependent variable. Drop the features that have high degree of collinearity with each other. As the number of features wsa small this was done manually. Similarly this could have been achieved using SelectKBest and PCA procedures and Recursive Feature Elimination (which was used to verify our features).
- For scaling the data with a skewed distribution it is important to use a distance-based classifier such as KNN and SVC.

Supervised Learning

- Training and Testing of dataset
- Classifiers: Logistic Regression, K-Nearest Neighbors, Support Vector Classifier, Elastic Net
- Model Evaluation Metric: Accuracy, Precision, Recall, F1, AUC

Classifiers	Precision	Recall	F1-score	Accuracy score	AUC
Logistic Regression	0.00	0.00	0.00	0.70	0.13
Logistic Regression Elastic Net with GridSearchCV	0.00	0.00	0.00	0.84	0.95
K-Nearest Neighbors (n=5)	0.82	0.79	0.81	0.86	0.90
K-Nearest Neighbors (n=7)	0.85	0.76	0.80	0.86	0.91
KNN with Standard Scaler	0.38	0.42	0.40	0.94	0.94
Support Vector	0.67	0.17	0.27	0.73	0.87

Conclusion

- For the purpose of this study testing scores were observed from classification report. Accuracy, precision, recall, f-1 score and AUC 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 with KNN with neighbors value of 7 closely following.
- SVC when used in conjunction with Standard Scaler had the strongest fit scores with an ideal AUC. The best AUC with proper overall fit was observed with KNN with Standard Scaler but had a significantly weaker F-1 score. Close secondary choice would be KNN with 7 neighbors.