Ruby Link

CPSC 392

May 2023

Variables

- ad_id: is the id of specific ad set, it is an numerical feature
- reporting_start: the date the ad was issued
- reporting_end: the date the ad was taken down
- xyz_compaign_id: is the id assigned by the ad running company
- fb_campaign_id: the id assigned by facebook for every ad set
- age: targeted age for the ad, in 4 year ranges
- gender: targeted gender of the ad
- interest: the user interests and likes of facebook users who were targeted for the ad (as provided in their bio)
- impressions: the number of times the ad was shown to the users
- clicks: number of clicks on the ad
- spent: amount of money paid to Facebook for that ad
- total_conversion: total number of people that signed up or installed the app/product
- approved_conversion: number of people that became actual active users

Q1

Question: (Supervised model) Which variable (spent, clicks, impressions) improves the model the most in terms of prediction accuracy for number of total conversions?

Variables Involved: spent (continuous), clicks (continuous), impressions (continuous), interest (continuous), total conversions (continuous), age (categorical), gender (categorical), campaign_id (categorical)

Cleaning: missing values will be dropped

Modeling/Computation: A Train/Test split with 80/20 split will be used and continuous variables will be z scored. We will use four linear regression models to predict total conversions. One model will use 8 of the 11 predictors listed above (the 11 minus the three we want to see the effect of- spent, clicks, and impressions). Next another linear regression model will be fit with the 8 predictors in addition to the variable clicks, another model will be fit with the 8 predictors but this time also including the variable spent. One last linear regression model will be fit with

the 8 predictors in addition to the variable impressions. R2 and accuracy scores will be calculated for train and test sets on all four models and compared.

Graphs: We will create a bar chart showing the train and test R2 scores for the four models. The predictor (spent, clicks or impressions) that improves the R2 the most will then be made into a scatter plot to show the relationship of it with total conversions.

Brief Discussion of why analysis is effective at answering the question: This analysis is effective because it compares a model that uses all but the predictors we want to analyze, with three separate models that each include one of the target predictors in order to see how adding these predictors improves R2 scores. The plots help us visualize the numeric results, as well as clearly demonstrate the relationship between the predictor that most improved prediction accuracy (spent, clicks, or impressions) and total conversions.

Q2

Question: (Clustering) How/do the clusters differ with the variables spent and approved_conversion for the three different campaigns?

Variables Involved: spent (continuous), approved_conversions(continuous), campaign_id (categorical)

Cleaning: missing values will be dropped.

Modeling/Computation: Continuous variables will be z-score. We will then use K-means to fit three different clustering models with the variables spent and approved_conversions, one for each of the three ad campaigns. Silhouette scores will be calculated for the three models and each will be compared.

Graphs: we will create a scatter plot for each campaign that colors the points based on cluster assignment.

Brief Discussion of why analysis is effective at answering questions: This analysis is effective because it allows us to compare what clusters form when looking at the variables spent and approved conversions but for each individual ad campaign as well as the characteristics of those clusters. This will give us insight into the relationship between spent and approved conversions but also how that differs based on the three different ad campaigns.

Q3

Question: (Dimensionality reduction) What are the most important features when predicting total_conversion for our ad campaigns using regularization models (ridge and lasso)?

Variables Involved: campaign_id (categorical), age (categorical), gender (categorical), interest (categorical), impressions (continuous), clicks (continuous), spent (continuous), approved_coversion (continuous)

Cleaning: To clean the data, we will drop the reporting_end_date because it is the same as the start date. We are also going to convert age to be categorical, sorting by each age range. We will also drop all of the null values and exclude the remaining variables not being used.

Modeling/Computation: We will have a train test split (80/20). Then fit the data to a lasso model and print out the train and the test MSE. Then, use the same split but on a PCA model and print out the MSE for the test and the train data. Compare the number of non zero coefficients for lasso and the number of PC's for PCA.

Graphs: Print out a cumulative variance and an explained variance plot to select the number of PC's. Also print out a bar chart of the coefficients.

Brief Discussion of why analysis is effective at answering questions: This is effective because a lasso model penalizes the sum of the squared coefficients, and will pull the values closer to zero. We can see which variables are most important by looking at the magnitude of the variables before and after the lasso model. The most important variables will have non-zero coefficients and the least important variables will either be dragged to zero (lasso).

Q4

Question: (Supervised model) Using a linear regression model, how does age (categorical) , gender (categorical), campaign_id (categorical), impressions (continuous), and spent (continuous) affect the clicks (continuous)?

Variables Involved: Age (categorical), gender (categorical), campaign_id (categorical), impressions (continuous), spent (continuous), clicks (continuous).

Cleaning: To clean the data, we will make a new column called ad_duration which is calculated by subtracting the reporting_start and the reporting_end. We are also going to convert age to be a categorical variable. We will also drop all of the null values and exclude the remaining variables not being used.

Modeling/Computation: We will begin by making a train-test-split (80/20) and z score the continuous variables. Then we will fit the data to a linear regression model. We will then use the train and test data to predict the number of clicks for our model and print out the MSE and R Squared value. We will also print out the coefficients.

Graphs: A bar chart showing the coefficients of all of the variables will visually show which predictors have positive and negative relationships with the number of clicks and which have the largest magnitude. Also, a bar graph of the R2 values for the test and the train data.

Brief Discussion of why analysis is effective at answering questions: By performing the linear regression and displaying the magnitude of the coefficients, we will be able to see how the different coefficients impact the number of clicks. It will show which ones have the largest effect with their magnitude and also the relationship (positive or negative) with the number of clicks and overall, which factors lead people to click or not click on an ad.

Q5

Question: (Clustering) When considering Spent and Impressions what clusters emerge and what characterizes these clusters? DBSCAN

Variables Involved: Spent(continuous), Impressions(continuous), and Clicks(continuous)

Cleaning: Missing values will be dropped

Modeling/Computation: Prepare the dataset, then choose appropriate values for the two key parameters for the DBSCAN. Use the elbow method to identify the best eps for the model. After fitting the model, analyze the clusters by printing out the silhouette score and scatter plots.

Graphs: Scatterplots colored by cluster assignment

Brief Discussion of why analysis is effective at answering questions:. DBSCAN is effective at answering this question because DBSCAN can identify noise points that do not belong to any cluster. This is important when analyzing real-world datasets that typically contain complex and irregular patterns; DBSCAN can detect these better than other clustering methods and our data may have some outliers which would make this form of clustering very useful.

Q6

Question: (Supervised) Using a decision (regression) tree to predict, which campaigns are most successful in getting the targeted audience engaging with ads?

Variables Involved: Campain_id (categorical), age (categorical), gender (categorical), clicks (continuous), total_conversion (continuous), impressions (continuous)

Cleaning: Missing values will be dropped, split data into training and test sets.

Modeling/Computation: Prep the data by dropping irrelevant and missing values, split the data into training and testing sets and create a decision tree model to predicate total_conversion. Fit the data to a decision tree model and output the MSE for the training and the test. Then create a new decision tree, this time setting min_samples_leaf and max_depth and calculate the MSE again to compare.

Graphs: A scatter chart would effectively depict the groupings of different input features predicting the target variable “total_conversion”, colored by each campaign. The x axis could represent the number of clicks and the y axis, the total_conversion.

Brief Discussion of why analysis is effective at answering questions: This analysis can provide insights into important audience segments, ad performance for each campaign, and enable data-driven decision making. Decision trees are fairly easy to interpret and the graphs will help visualize how the model is making splits.