Work on Campaign Forecasting under the Guidance of Prof B. Chandra.

Present Status/Discussion points for next meeting:

1. Working with Page Types type features
2. Retrieving ad level broken down data for country at AD Level. Presently we retrieve only for the Ad Set Level
3. Results of tested models: Linear Regression, GRNN.

Day wise work Log.

21st Nov

Ideation level. We want a model that extracts important features, trains the model and reports the result on the test set.

Thoughts on the project were:

- what all comprises the project?
    - are we predicting at the PI level or at the Ad Level?
    - challenge either way as the budget and the bidding will be a problem.
        - the same budget and the bidding are shared by different ads/ PIs
        - so different start/end times in adSet/PI
        - this will create a problem - total_budget_amount_local_micro – means different budgets at adSet and PI.
        - expect issues when working with target CPA and optimization for website conversion.
        - Might need a separate model for this one ^
        - How to take into account the metrics that are additionally available. For example for a tweet using Tweet Engagements, there is a Video available and hence the video metric is not null which representative of the outcome.
        - What is the data we can use? Can we only use account restricted data? Can we use cross partner data? Can we use global data?
        - We could use the partner level and global data with representative weights to symbolize the distance in the data.
        - How can the user visualize the model?
        - what to optimize on? Have Coke made any requests for any particular metric.
        - Future possibilities: http://www.optimove.com/learning-center/machine-learning
    - **assume for now all are independent.**
    - from the adset:
        - the targeting. some more important than others. like the **interest** and the country and **keyword** probably. maybe like events
        - adset budget as well, if we support.
        - the bid type
        - the **bidding**
            - bid value, type, goal, billing
    - from the PI:
        - the lifetime and daily budget.

Work at prototype

We extracted sample data from Coke environment, across accounts. We trained linear regression models in MATLAB.

Mean squared error = 4.00354306891e-06

22-23-24: leave. No leaves during Diwali/Dussehra to take leaves during this time of year.

25th Nov:

Extracting columns from the sample data. Initial columns for training.

Ad Delivery Status
Automatically Set Bid
Bid Type
Budget Pacing
Max Bid

Paid Initiative Daily Budget – Removed later after discussion with Manish sir. Interested in rate (ex CPE)
Paid Initiative Lifetime Budget – same
Clicks/Impressions

Facing the "scale" problem in MSE, we migrated to using R-Squared as the error statistic. R-squared is a metric used for testing a model's prediction ability. It calculates the ratio of variance captured by the model against the variance in the original data.

We trained the model using non linear relations using the MATLAB code :

```
link = @(mu) log(mu ./ (1-mu));
derlink = @(mu) 1 ./ (mu .* (1-mu));
invlink = @(resp) 1 ./ (1 + exp(-resp));
F = {link, derlink, invlink};
b = glmfit(X,Y,'binomial','link',F);
```

The models trained using made some valid predictions for the first time. They are attached.

Best R-Squared result - 0.66
MSE - 3.56E-06

Studied about Logistic Regression and discussed with Ma'am about how it was not appropriate for our model.

28th Nov:

Worked on concrete steps laid out by Prof B. Chandra:

1.) Divide the training and test data and find the R squared value for the test data

Procedure: I randomly distribute values using the numpy random function and choose approx. 70% of the data to train. It could be 69% and could be 71%

Training set R - Squared Value -
0.542131609113
Test Set R Squared Value -
2.63777812857 - prediction is inconsistent in the middle of the plot.

I have attached the plots titled "28thNov_TwoInstances".
In each plot, the actual values of conversion for the ad instance are in Dotted Blue and the predicted values are in Red. For example: "networkresultForTestCase_93UnseenCases" shows the prediction on the UNSEEN test data set (30% of total model) in red.

2.) Include more relevant features and see which ones' increase R-Squared value.
I have added Gender. The plots are attached.
Note : There were only 7 instances of not null values in the 313 cases. In the other Ads, no gender was selected.
Results on adding Gender – 1st Training
Training set R - Squared Value - 0.874335935164
Test - 0.0723949047751
Results on adding Gender – 2nd attempt:
Training set R - Squared Value - 0.559553912263
Test - 0.915784360749

The 2nd attempt is much better that the 1st as the Models tend to get in local Minima. To overcome this, we decided to use ten cross ten validation.

29th Nov:

Created a cross validation function using the MSE cost function, error = 4.4268e-06.
R Squared Value and 10-fold cross validation gives average = 11.6002, results:

- 3.1360
- 0.5103

- 0.0409
- 8.0966
- 0.2232
- 72.9807
- 23.6814
- 0.0543
- 5.2703
- 2.0086

Here we are using the ESS/TSS model. The R-Squared value above 1 represents poor results.
I have attached each of the ten test predictions in the above.
BusinessLocationRefreshJob issue in the evening.

30th: leave

1st Dec:

- The linear regression with non linear relations was put aside to experiment with GRNN.
- In case of categorical attributes, give probability of occurrence of those values. Example, if there are are 7 yes and 3 No, 0.7 to be given to all yes and 0.3 to be given to all nos.
- Studied Radial Basis Functions. GRNN is a Neural Network based model with RBFs driving the activation of a single hidden layer in the network.
- They have a value called Spread, which determines the sensitivity of the model. We experimented with Spread values 0.2 to 0.9 and the resultant plots are attached.
- We experimented with:
  - removal of lower/higher thresholds for the conversion rate.
  - Different default values for better prediction
- Result Instance: for Spread - 0.4
  - R-Squared Train = 0.0869
  - MSE Train = 2.9328e-07
  - R-Squared Test = 0.0797
  - MSE Test = 5.5492e-07
- BusinessLocationRefreshJob issue in the evening – Restricted API in the evening.

2nd Dec

1.Trying to Run General Regression Neural Networks with categorical values rather than probability for spread 0.4, 0.6, 0.8. In the images attached, the spread is in the name, the R-Squared value and the MSE are in the name. For example in the name: "Train_Plot_Spread_0.4_rsquared_0.4114_mse_4.9601e-06.png", spread == 0.4, R-Squared == 0.4114 and MSE = 4.9 e-06

2. Read and try Probabilistic NN

PNNs can find decision boundaries and are used for Classification. GRNN are used for regression.

3. Use some other measure for bringing categorical values between 0 and 1
I used the summation of probabilities (https://en.wikipedia.org/wiki/Cumulative_distribution_function). I sorted the keys based on the probabilities in the ascending order and for each category:

```
for tuple in probability_set:
    summation += enum_probability_value
    probabilities_dict[tuple[0]] = summation (edited)
```

Why Cumulative Distribution :
Goal - map categorical values to value between 0,1

But, using just Probability: Distinction is lost!

Initial Data	Processed
A	0.5
A	0.5
B	0.5
B	0.5

Using Cumulative Distribution: Uniqueness retained

Initial	Processed
A	0.5
A	0.5
B	1.0

B 1.0

4. Include all features and try all these techniques.

Following were added:

- Countries, in our data each row only corresponds to one or less countries.
- Billing Event
- Optimization Goal
- Frequency Control – Frequency Cap and Duration Days
- For Page types, Keywords, Events, I will need to discuss the approach.