InPowered
Data Challange

Proposed Solution

The main idea of the proposed solution relies on training Machine Learning (ML) models to minimize the Cost Per Engagement (CPE) for a particular advertising item, which comprises different targeting, creative (headline/summary) and CPE values.

There available data is composed of categorical and numerical values and can be found on the *problem_merged_data.csv* file.

The idea of minimizing CPE derives naturally from its definition. Given a target item, the CPE can be described as

$$\text{CPE}\ (item) = \frac{\text{Media Spend}\ (item)}{\text{Engagement}\ (item)}$$

Ideally, it would be possible to diminish the cost (Media Spend) and increase the Engagement simultaneously. This is not always true, such that the proposed approach goal is to **find the best Bid x Budget values that minimize CPE** for a particular *item*.

To this end, the focus of this strategy is to train two estimators, one for each metric adopted in the CPE, i.e., the estimators $\hat{y}_{\text{spend}}$ and $\hat{y}_{\text{engagement}}$.

Assumption #1:
Media Spend and Engagements are **not** necessarily functions of the same set of variables.

Although both target metrics are most likely functions of the available multidimensional data, the set of values which directly impact each case may differ.

Therefore, it is defined the independency assumption over each estimator. Moreover, given the estimated values of **Media Spend** and **Engagement**, the estimate **CPE** (to be minimized) is defined as

$$\hat{y}_{\text{CPE}}\ (Bid, Budget) = \frac{\hat{y}_{\text{Media Spend}}\ (Bid, Budget, Target_s)}{\hat{y}_{\text{Engagement}}\ (Bid, Budget, Target_e)}.$$

Data Inspection

As means to corroborate the **Assumption #1,** the data inspection was carried for both categorical and numerical data. In this report, the most important insights are highlighted for each case considering the target metrics of **Media Spend** and **Engagement**.

CATEGORICAL DATA

The first inspection on this type was carried based on the number of unique values in each data-column available in the dataset (1682 items total):

DATA	# UNIQUE VALUES
ITEM	157
CHANNEL	3
DATE	358 (7 WEEKDAYS)
HEADLINE	68
STORYSUMMARY	35
IABCATEGORY	4
CATEGORY_1	8

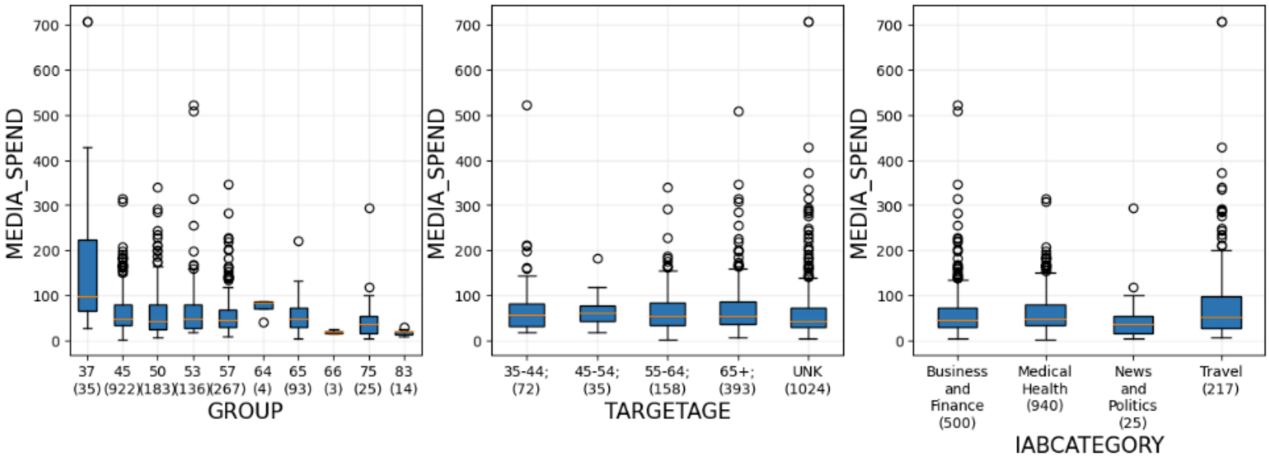
DATA	# UNIQUE VALUES
TARGETGEO	16
TARGETINTEREST	25
TARGETAGE	5
TARGETOS	3
TARGETDEVICES	2
TARGETGENDER	3
TARGETLANGUAGES	2

Insights #1:

- **DATE:** 358 days reduced to 7 weekdays - Media Spend and Engagement might vary over different days, for instance, it might occur that on weekends the engagement is higher.
- **CHANNEL:** In the available data there is only one instance of MGID channel, for modeling and estimation it might be interest to acquire more examples of this case.
- **TARGETGEO** and **TARGETINTEREST**: Although these are filled with multiple dictionaries under each item with considerable raw data, there are few unique examples in each case (16 and 25, respectively). This indicates that an embedding approach could be considered for the feature representation of theses topics.
- **HEADLINE** and **STORYSUMMARY**: These are the most semantic-based (formal, lexical and conceptual) data available. The combination of Headline and Story Summary may significantly impact the customer's behavior by inducing or instigating a desired action. Sentiment analysis feature representation of this data might be desired.

The inspection of each categorical data was also carried out by analysis of the distribution using Box-Plot tools and Tukey HSD Pairwise Group Comparison with 95% confidence interval.

As an example, for **Media Spend**, it is noticeable that the Group is influential over the resultant metric, whereas TargetAge known cases do not seem to have any influence and IABCategory have mixed similarity distributions over each of the four cases.



The known TargetAge p-values are all above 0.843 which implies that its distributions are not statistically different for the **Media Spend** evaluation, statistical difference occurs for p-values below 0.050.

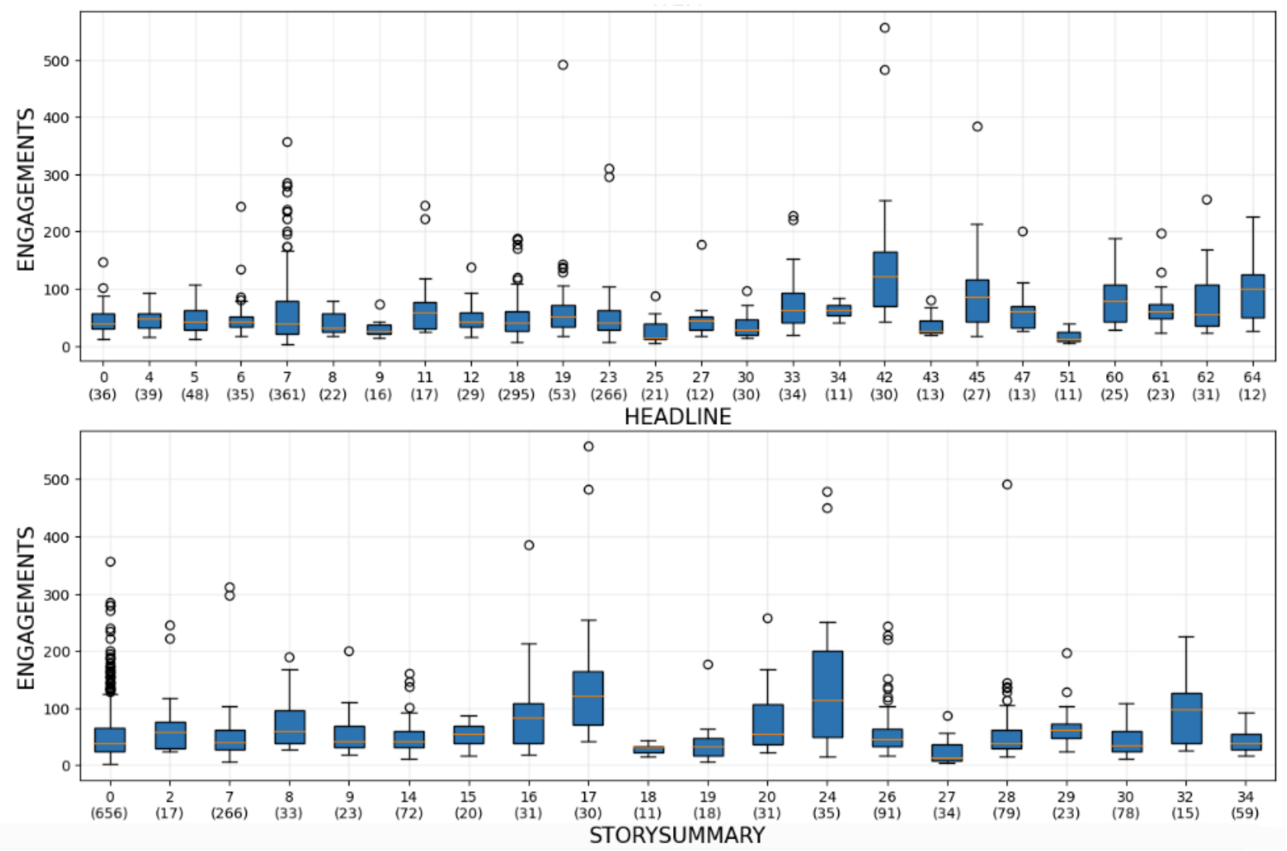
To illustrate this behavior, HSD Pairwise Tukey evaluation for the IABCategory is fully presented below:

IABCATEGORY	
Pairwise Set	**P-Value**
(0-1)	0,983
(0-2)	0,760
(0-3)	0,000
(1-2)	0,809
(1-3)	0,000
(2-3)	0,014

The data indicates that Travel category is the one statistically different than all the other groups, with maximum p-value of 0.014 < 0.050.

A similar analysis was performed considering the **Engagement** metric. In this case, there are also non-relevant categories such as Target Gender and Category_1. The influence of Item, Headline and Story Summary is consistent over this measure.

The last presented categorical data box-plot in this report is illustrated next for the Headline and Story Summary influence over the **Engagement** measure. In the plot, unique texts are replaced by numbers and it considers only distributions with over 10 examples for clarity:



The same behavior of Headline and Story Summary is also noticeable over **Media Spend** and consequently on **CPE**.

One main difference on the pairwise Tukey analysis for **Media Spend** and **Engagement** is present in the Category_1 data. In this scenario, the p-values of **Media Spend** are ranged between [0.000, 0.999] with mixed distribution similarity over the data. That does not occur for **Engagement,** which achieves p-values on the [0.361,1.000] range (all above 0.050).

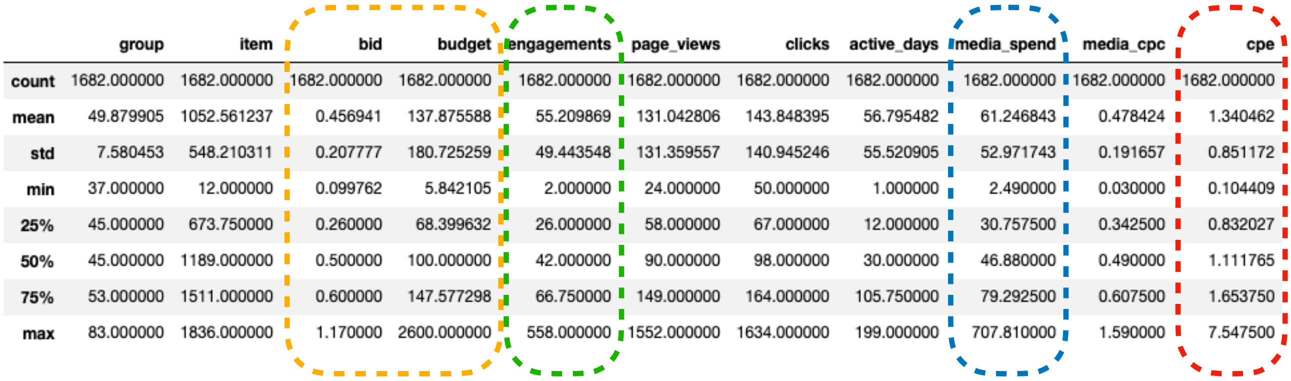
This implies that Category_1 is most likely to benefit the learning of the $\hat{y}_{\text{spend}}$ estimators and **not** $\hat{y}_{\text{engagement}}$. Which also reinforces the considerations presented in **Assumption #1.**

The idea of this subsection is to illustrate some of the most important insights over the available categorical data. A fully description and visualization can be assessed at the “Data_Visualization.ipynb” notebook.

For example, the Tukey evaluation of the **Date** x **Media Spend** and **Date** x **Engagement** indicates that this particular data has no influence whatsoever in the distribution outcome of the target measures and therefore might be omitted as model inputs.

NUMERICAL DATA

The numerical evaluation of the dataset was focused on the distributions, statistics and correlation over the data. A description of the data frame is presented below:



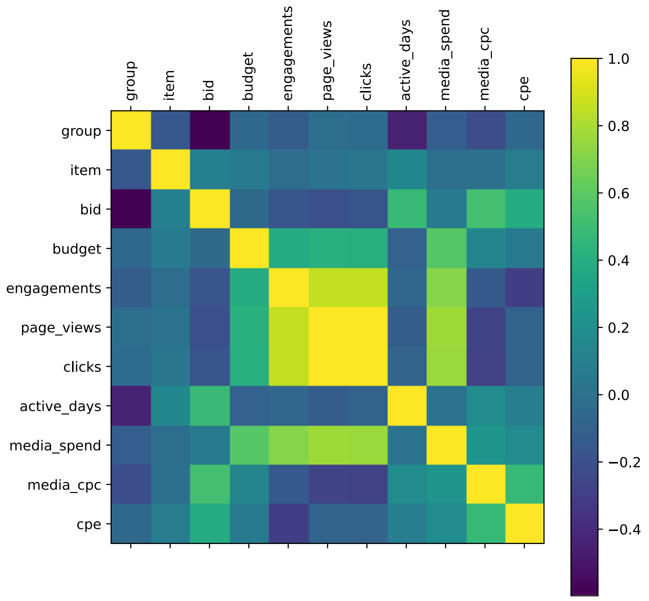
	group	item	bid	budget	engagements	page_views	clicks	active_days	media_spend	media_cpc	cpe
count	1682.000000	1682.000000	1682.000000	1682.000000	1682.000000	1682.000000	1682.000000	1682.000000	1682.000000	1682.000000	1682.000000
mean	49.879905	1052.561237	0.456941	137.875588	55.209869	131.042806	143.848395	56.795482	61.246843	0.478424	1.340462
std	7.580453	548.210311	0.207777	180.725259	49.443548	131.359557	140.945246	55.520905	52.971743	0.191657	0.851172
min	37.000000	12.000000	0.099762	5.842105	2.000000	24.000000	50.000000	1.000000	2.490000	0.030000	0.104409
25%	45.000000	673.750000	0.260000	68.399632	26.000000	58.000000	67.000000	12.000000	30.757500	0.342500	0.832027
50%	45.000000	1189.000000	0.500000	100.000000	42.000000	90.000000	98.000000	30.000000	46.880000	0.490000	1.111765
75%	53.000000	1511.000000	0.600000	147.577298	66.750000	149.000000	164.000000	105.750000	79.292500	0.607500	1.653750
max	83.000000	1836.000000	1.170000	2600.000000	558.000000	1552.000000	1634.000000	199.000000	707.810000	1.590000	7.547500

The solution goal is to be able to correctly predict **Media Spend** and **Engagements** ($\hat{y}_{\text{spend}}$ and $\hat{y}_{\text{engagement}}$) for a particular item, given a set of Bid and Budget, so that it is also possible to estimate the minimum **CPE**. These metrics are highlighted in the image above.

Insights #2:

- **Outliers:** Target measures present most of their data (75%) inside the $\mu + \sigma$ threshold. Nonetheless, checking the maximums in each case, it is evident the existence of outliers, values that are considerably above the $\mu + 2\sigma$ *or even* $\mu + 3\sigma$*.* This might impact the learning process of our estimators.

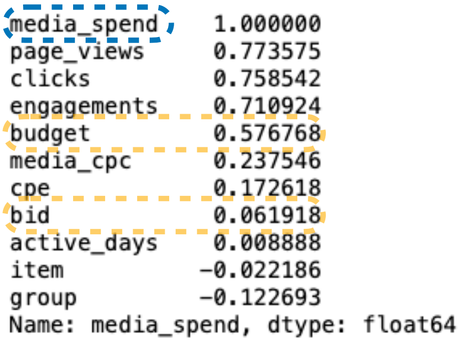
Another important aspect of the numeric data is how these distributions correlate with one another.



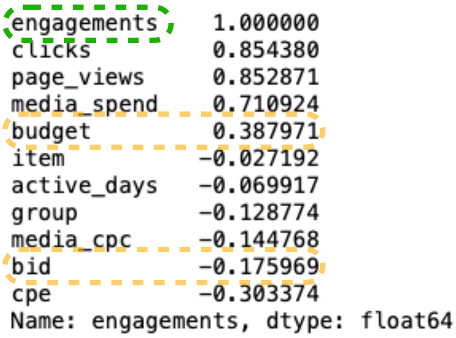
The correlation (ρ) of the numeric data is presented as a matrix image plot. The highest values of ρ occur between the following measures:

- Budget
- Engagement
- Page Views
- Clicks
- Media Spend

Note that Bid and Budget have different correlation values across the other measures, specially for targets **Media Spend** and **Engagement**. It would be interesting to take a closer inspection on the Bid and Budget relation regarding these measures. It is presented below the ρ sorted values for **Media Spend** and **Engagement** considering all the other numeric data distributions available.



```
media_spend      1.000000
page_views       0.773575
clicks           0.758542
engagements      0.710924
budget           0.576768
media_cpc        0.237546
cpe              0.172618
bid              0.061918
active_days      0.008888
item            -0.022186
group           -0.122693
Name: media_spend, dtype: float64
```



```
engagements      1.000000
clicks           0.854380
page_views       0.852871
media_spend      0.710924
budget           0.387971
item            -0.027192
active_days     -0.069917
group           -0.128774
media_cpc       -0.144768
bid             -0.175969
cpe             -0.303374
Name: engagements, dtype: float64
```

Note that Bid and Budget present different ρ values over target measures. Budget is positive correlated in both cases with stronger correlation in **Media Spend** (0.58 over 0.39). On the other hand, Bid is actually negative correlated to **Engagement** (-0.18) and has a small positive correlation with **Media Spend** (0.06).

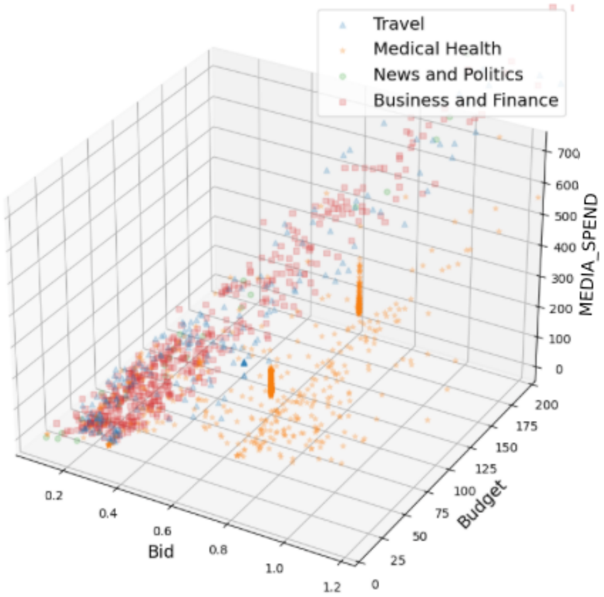
Insights #3:

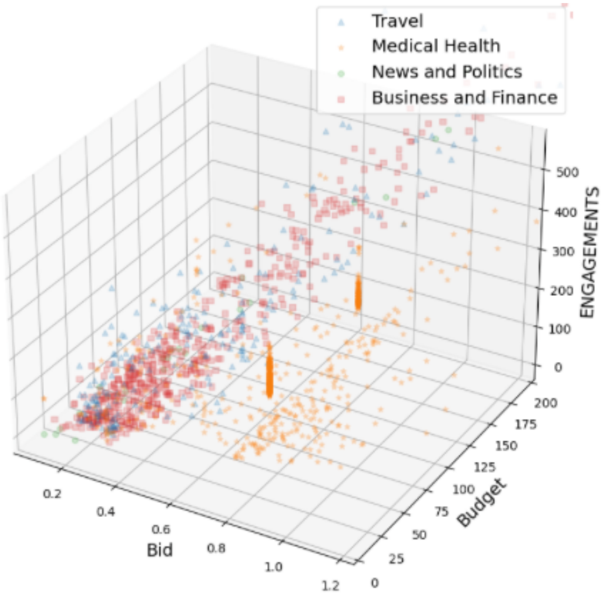
- **Budget x Spend:** There is a strong correlation between the Budget amount and the Media Spend. This is a solid evidence that, for our dataset,

  "If someone is willing to pay more for something (Budget) it is more likely than not that he/she will pay more (Media Spend)."

- **CPE Minimization:** The small and negative ρ values observed for Bid, in conjunction to the Budget observed correlations, indicate that it is likely to exist a trade-off on the Bid-Budget pairs such that Media Spend and Engagement can be adjusted to minimize the **CPE**.

With these insights in perspective, it would be compelling to evaluate the occurrence of values in the Bid x Budget x Measure space.





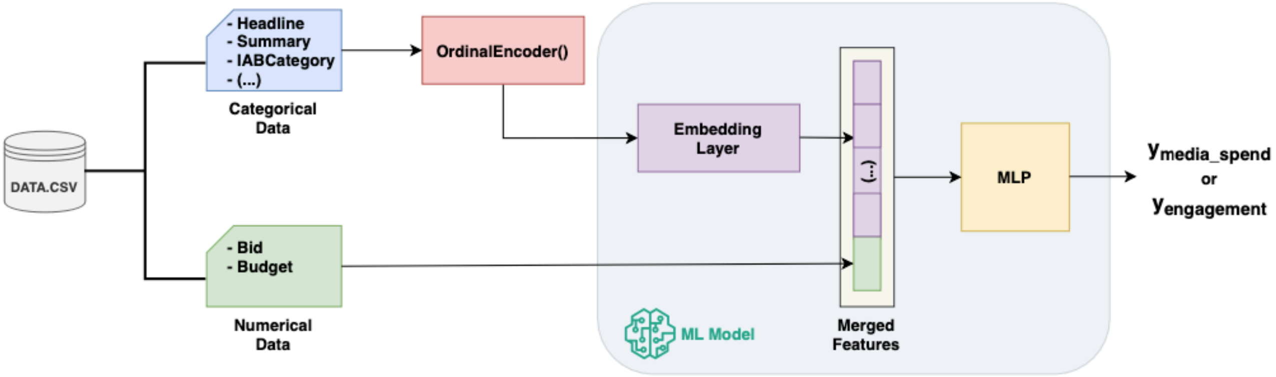
In these scenarios, it can be observed the positive numeric correlation of the target measures regarding the Budget variable. Conversely, this is not the same trend for Bid., as expected.

Insights #4:

- **Data Quirks:** One potential issue illustrated in this scatter plot is the existence of data quirks (vertical lines in clear Bid x Budget) regions. For a final solution, it would be desired to remove these data quirks to avoid the algorithm to learn this specific quirks.

Baseline Solution

The baseline solution is developed considering all the available data divided into Category and Numerical as presented below:

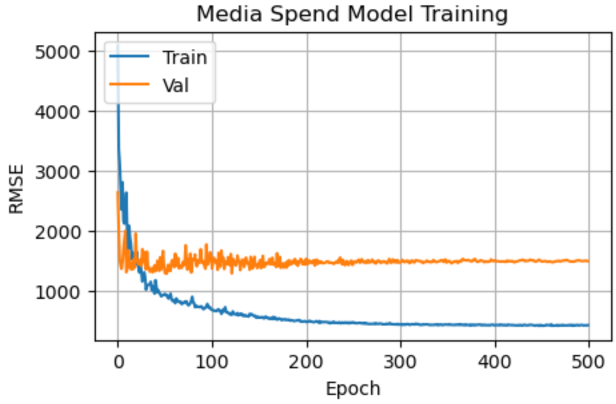


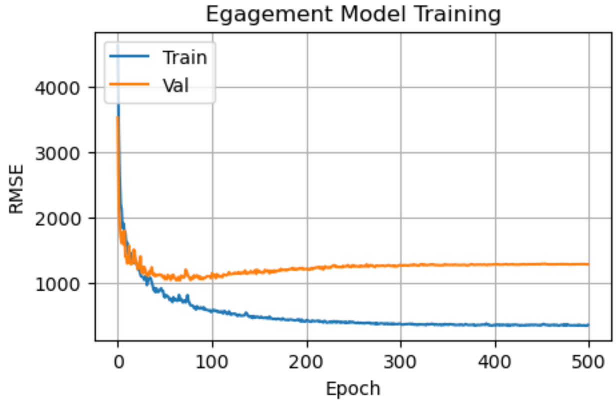
Assumption #2:
The small number of categorical unique values can be represented by embeddings.

The baseline solution adopts an embedding representation for each categorical feature. The **Assumption #2** encodes the unique values of each particular data of the dataset into a vectorial embedding representation. During the training step, the model will learn to represent each possible categorical value into a multidimensional space. Related data will be comprised into similar regions, most likely distant from non related ones.

The ML model adopted for this solution receives as inputs the ordinal encoded values of categorical data and the (Bid, Budget) numerical data for a particular item. After the learnable embedding layer a Merged Feature is created, concatenating the (Bid, Budget) pair to all embedding representations which are them passed into a Multi-Layer Perceptron (MLP) with single dense output and linear activation.

The MLP is composed of six dense layers of 1024 units followed by a 16 dimensional dense the the output. The same model architecture is adopted for both $\hat{y}_{\text{spend}}$ and $\hat{y}_{\text{engagement}}$ estimators.

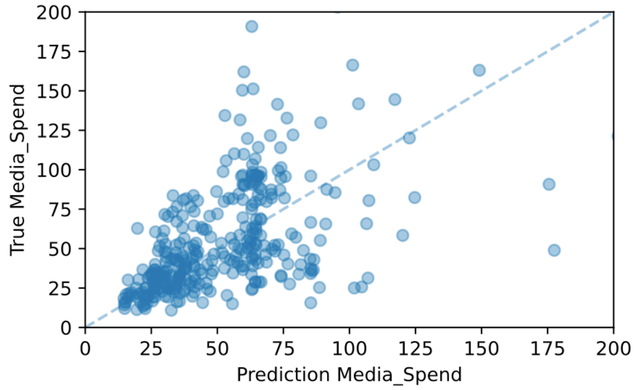


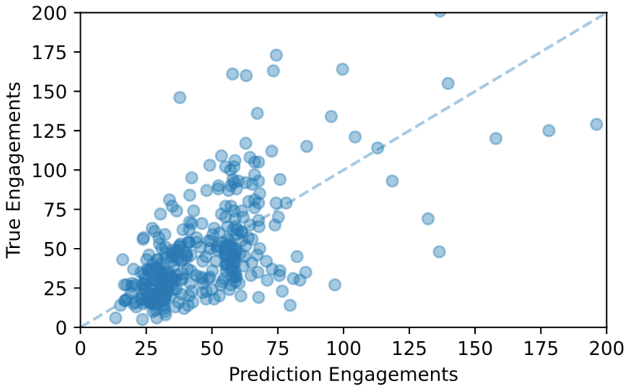


The training is performed considering a 80%/20% data split with Adam optimizer and the RMSE loss for 500 epochs. Validation and Test set are kept equal. The Models converge better as a Engagement estimator than Media Spend, and starts to overfit close to the 50th and 20th epochs, respectively. More epochs and higher decay rates may lead to lower RMSE values.

This behavior (variance issue) is expected due to the amount available data, and possible solutions relies on the usage of more data, synthetic data augmentation or other model architectures.

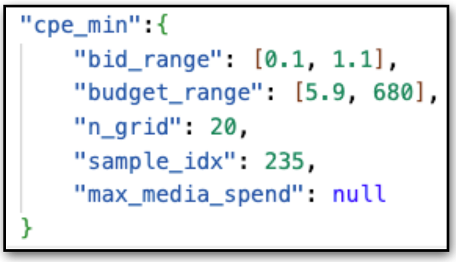
After training, each model is used to estimate $\hat{y}_{\text{spend}}$ and $\hat{y}_{\text{engagement}}$ in order to acquire the minimum $\hat{y}_{\text{CPE}}$.





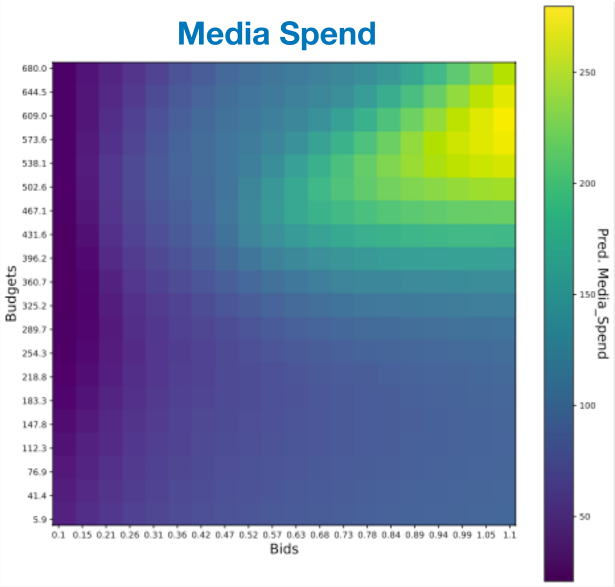
The Predicted x True values for both estimators in the test dataset reinforces the model convergence. Ideally all scatter points would be located in the dashed line. There is a clear and logic organization of the dots around the line, specially for values that are close to the distribution mean. Values tend to be more *off-the-line* for true values that present outlier behavior.

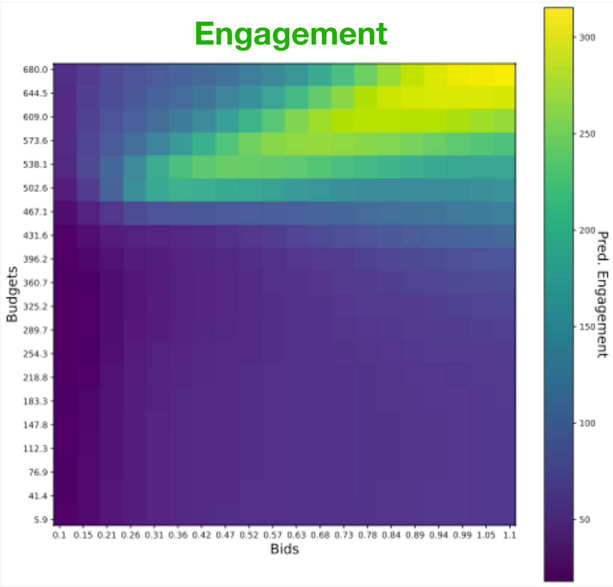
To minimize the CPE, it is than considered for each item a grid Bid x Budget which is comprised by default in [0.1, 1.1] x [5.9, 680] as presented in the "*cpe_min*" values of the JSON config values of this solution. The existence of a maximum desired Media Spend is also considered for the evaluation.

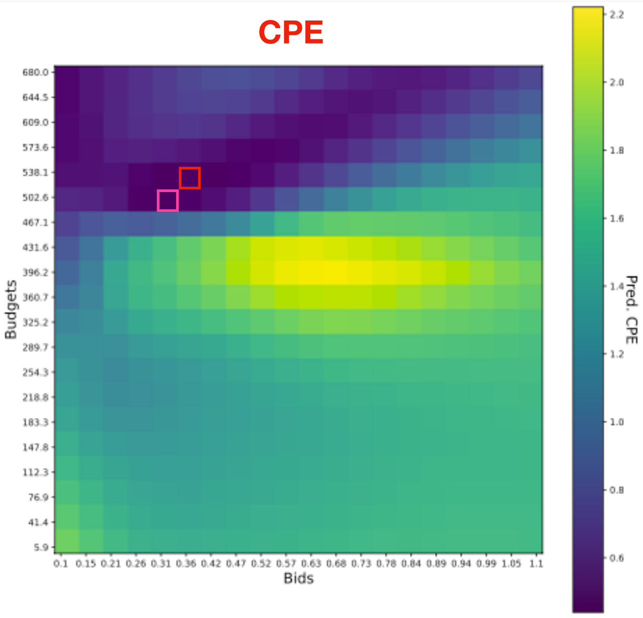


```
"cpe_min":{
    "bid_range": [0.1, 1.1],
    "budget_range": [5.9, 680],
    "n_grid": 20,
    "sample_idx": 235,
    "max_media_spend": null
}
```

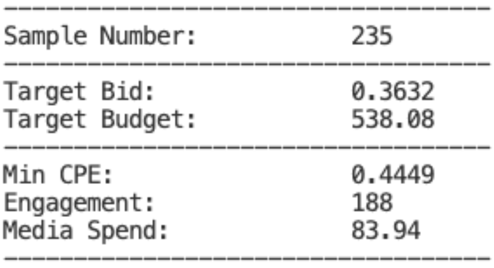
Next, it is present the predicted Media Spend, Engagement and CPE Grids for this solution and a sample output of the approach developed.







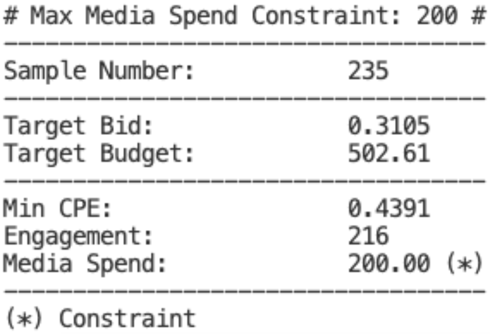
The resultant $\hat{y}_{\text{CPE}}$ is computed as the division of $\hat{y}_{\text{spend}}$ and $\hat{y}_{\text{engagement}}$ estimators in the grid. The red square represents the minimum CPE value obtained in the estimated Bid x Budget grid.



```
------------------------------------
Sample Number:              235
------------------------------------
Target Bid:                 0.3632
Target Budget:              538.08
------------------------------------
Min CPE:                    0.4449
Engagement:                 188
Media Spend:                83.94
------------------------------------
```

Bid: 0.3632 | Budget: 538.08

The pink square considers a constraint of maximum Media Spend of U$ 200. Note that in this case, both Bid and Budget values are changed to a new minimum **CPE** that also considers the spend constraint. **Bid: 0.3105 | Budget: 502.61**



```
# Max Media Spend Constraint: 200 #
------------------------------------
Sample Number:              235
------------------------------------
Target Bid:                 0.3105
Target Budget:              502.61
------------------------------------
Min CPE:                    0.4391
Engagement:                 216
Media Spend:                200.00 (*)
------------------------------------
(*) Constraint
```

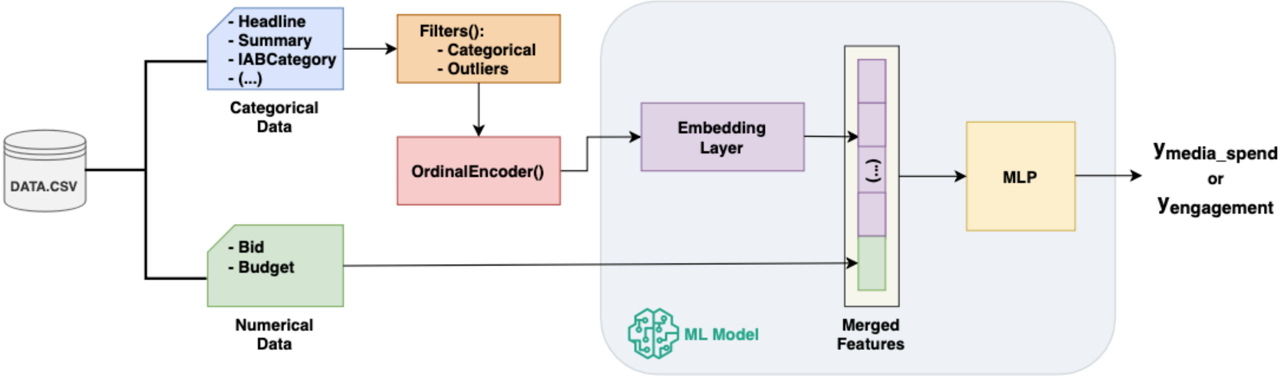
Meaningful Inputs

As further considerations regarding the proposed solution, and based on the data insights carried before it would be reasonable to consider assumptions to improve the current approach.

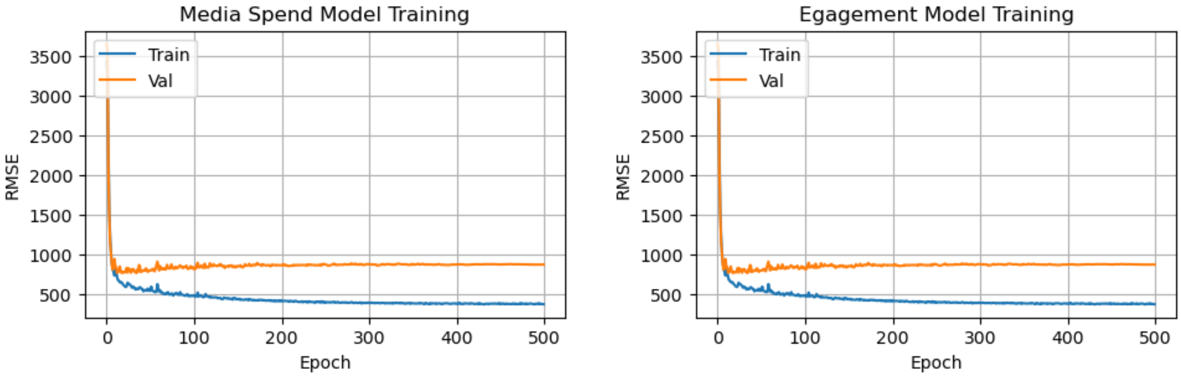
Assumption #3:
Meaningful categorical data and outlier removal will improve model convergence.

In the following case, the **Media Spend** and **Engagement** values x that are considered outliers, i.e. $x > \mu + 3\sigma$, are filtered out (59 samples, only 3.5% of the data).

Moreover, exclusively the categorical data that were considered statistical significant over the Tukey's HSD Pairwise Group comparison are kept as inputs. This means that the input data presented at least one pairwise comparison which was significant over the Turkey statistical evaluation.



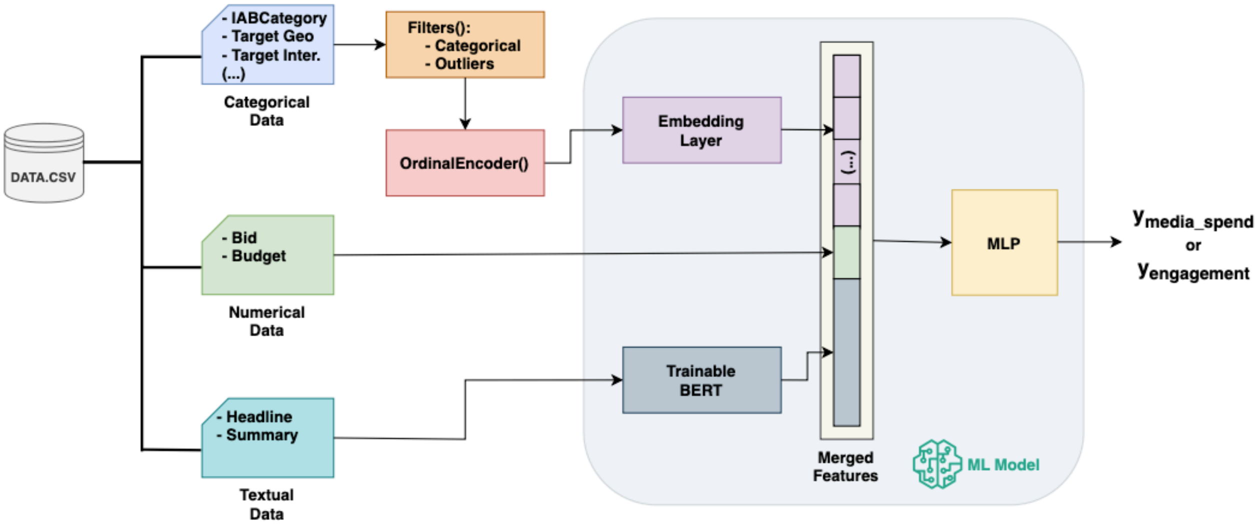
In this case, the ordinal encoder considers different categorical data for **Media Spend** and **Engagement**, due to the filtering step. The model difference relies on the number of learnable categorical embedding features available per model.



Model convergence is significantly improved, specially for the **Media Spend** estimator which is less noisy in the validation set. Both training losses reach smaller values with this adjustment.

The BERT-MLP

The final consideration regarding the proposed solution relies on the Headline and Summary categorial / textual data available in the dataset and also discussed in the **Insights #1**.



Headline and Summary are semantic-based textual data. The combination of Headline and Story Summary may significantly impact the customer's behavior by inducing or instigating a desired action.

Therefore, sentiment analysis feature representation of this data might be desired. The Bidirectional Encoder Representations from Transformers (BERT) is a pre-trained model which have been successfully adopted in a variety of of textual and language applications and could be applied as trainable attention layers inside our solution to acquire representative features for the Headline+Summary composition.

In this modification the textual semantic data is now included as a model input which will be considered by the BERT module to create representative features. The next steps remain the same, the Merged Feature is passed to the MLP layers in order to have have the predictions of $\hat{y}_{\text{spend}}$ and $\hat{y}_{\text{engagement}}$.

The resulting BERT-MLP model is considerably bigger (35M params x 5M params) than latter models. In the repository shared with this report, the "models/bert_mlp.py" should be replace by the code of "models/real_bert_mlp.py" in order to perform this task. A Google Colab adaptation of the code "BERTonColab.ipynb" was also shared to illustrate this stage of the solution, although some adaptations on paths and variables is necessary. The implementation of these results can be found at:

https://github.com/g-zucatelli/inpowered/

Final Considerations

There are some aspects of the solution that could lead to better models and predictions. As far as possible next steps go, in the investigation it could be interesting to:

- Completely train the BERT-MLP model and evaluate its predictions against the current solution.

- Train the MLP-based models in a bigger dataset and check if the variance problem noticed during training is solved.

- Evaluate if a classical word2vec compositions over the words of Headline and Summary data could improve the model.

- If there are not more data available, perform data augmentation: utilize statistical models (such as GMMs) to fit the multidimensional data and draw new samples from the distribution.