Team 3:

- Agarwal, Ayush Gordhan
- Gonzalez, Camila
- Rao, Abhishek
- Lu, Jianyuan

Date - 12/09/2021

Customer Data for Advertising Campaigns

ISYE 6414 – Final Project

Table of Contents

I. Problem Statement

Advertisements in social media have become more popular with the continuous increase in number of users in platforms like Facebook, Instagram, and YouTube. Can we predict whether Abhishek will purchase a product based on whether he clicked on the ad that popped in his Instagram feed?

With this project, we aim to understand what factors influence customers' spending behavior. To accomplish this, we will attempt to predict customers' spending based on their exposure to Facebook Advertisements.

II. Data Sources

1. Dataset

1.1. Predictor Variables

The dataset[1] used recorded the interactions of more than 1,000 Facebook users with different advertisement campaigns. It contains both information about the user itself (age, gender, interests) and about the advertisement the user is exposed to. The list below defines the predictor variables in the model.

- *age:* age of the user exposed to the ad.
- *gender:* gender of the user exposed to the ad.
- *interest:* represents the category to which the user's interests belong to based on the user's Facebook's profile.
- *Impressions*: represents the number of times the ad was shown to the user.
- *Clicks:* represents the number of times the user clicked on the ad shown.
- *Spent:* represents the amount paid by Company XYZ to Facebook to show their ad.
- *Total Conversion:* represents the number of people who researched the product after seeing the ad.

With all these variables, we will attempt to predict the response variable *Approved Conversion*, which represents the total number of people that purchased the product after seeing the ad.

In the given data set, the feature *interest* takes distinct values in the range of 1-120. However, the meaning of each interest value is not given in the data set. We cannot consider this feature as quantitative since each number has its own feature assigned. Also, considering *interest* as a categorical variable is not practical because if we do so our model would have 120 coefficients corresponding to each interest value. Hence, we assume that there are 5 categories of interest: 0-25, 25-50, 50-75, 75-100, 100-125, such that each interest category might contain the value of a single domain. E.g., If we assume that *interest* 1-25 is Sports then it might have sub-domains like 1=Football, 2=Basketball etc.

age	gender	interest	Impressions	Clicks	Spent	Total_Conversion	Approved_Conversion
30-34	M	0-25	7350	1	1.43	2	1
30-34	M	0-25	17861	2	1.82	2	0
30-34	M	0-25	693	0	0.00	1	0
30-34	M	25-50	4259	1	1.25	1	0
30-34	M	25-50	4133	1	1.29	1	1

Figure 1 – Screenshot of Dataset

1.2. Response Variables

In this study, we will fit two different models: a Multiple Linear Regression (MLR) and a Poisson Regression (PR). Since both models are, in general, made to fit two different types of problems, the response variables were adapted accordingly. With the MLR, we will predict the *Approved Conversion*, and with the PR, we will predict the *Click-Through Rate*.

Approved Conversion is defined as the total number of people who bought the product after seeing the ad. This is an important metric for the company because it would indicate the number of people that buy the product as a result of the ad campaign.

To verify the accuracy of the dataset, we validated the zero values for *Spent* and for *Approved Conversion.* Does it make sense to have in our dataset a company that spent $0 in a Facebook Ad campaign? We researched the policies of Facebook's ad campaigns, and it turns out Facebook does not require an initial investment to start showing the investor's ads in their platform and can charge per *Click* per *Impressions*.

Is it possible to have companies that invested in ad campaigns on Facebook, but did not receive any customers out of it? About 50% of the ad campaigns in the data have 0's for *Approved Conversion*. We can suspect that for about half of the users, purchasing from an online ad is not in their common practice.

The *Clicks-Through Rate (CTR)* is defined as the percentage of people who clicked an ad from the total number of people to whom the advertisement was shown (impressions). It is the ratio of clicks upon impressions, or, in brief:

$$CTR = \frac{Clicks}{Impressions} \times 100$$

In fact, *CTR* is a significant metric in the world of advertising because it helps an organization understand their target audience. A low *CTR* could indicate that a company is failing to speak the language of its potential customers. It could also indicate that the ad is not good enough to make them click or it does not coincide with their interest or several other reasons as well.

2. Exploratory Data Analysis

2.1. Data Analysis Overview

Statistics about the variables in the dataset are represented in Table 1 below. The values of the means for *Impressions*, *Clicks* and *Spent* are larger than the values of the medians. With this information, we can suspect that there is a cluster of large values for those predictors. To put in in context, we believe that companies either make small to medium-sized investments in ad campaigns, or very big investments.

Variable	Min	1st Qu	Median	Mean	3rd Qu	Max
Impressions	87	6504	51,509	186,732	221,769	3,052,003
Clicks	0	1	8	33.39	37.5	421
Spent	0	1.48	12.37	51.36	60.02	639.95
Total Conversion	0	1	1	2.856	3	60
Approved Conversion	0	0	1	0.944	1	21

Table 1 – Statistics about Predictor and Response Variables

To further validate the accuracy of the data, we verified if any null values were present in any of the columns in the data. Luckily, there were no null values.

2.2. Correlation Analysis

By the correlation matrix displayed in Table 2, we can observe values ranging from 0.56 to 0.86. This is translated into the presence of a strong linear relationship between predictors themselves and with the response variable. This also indicates that the model may have multicollinearity issues.

	Impressions	*Clicks*	*Spent*	*Total Conversion*	*Approved Conversion*
Impressions	1	0.95	0.97	0.81	0.68
Clicks	0.95	1	0.99	0.69	0.56
Spent	0.97	0.99	1	0.73	0.59
Total Conversion	0.81	0.69	0.73	1	0.86
Approved Conversion	0.68	0.56	0.59	0.86	1

Table 2 – Correlation Matrix for Predictor and Response Variables

2.3. Visual Analytics

The graphs below show the relationships between each predicting variable and *Approved Conversion* for Multiple Linear Regression, and *CTR* for Poisson Regression.

Plots like those in Figure 2 were used to evaluate relationships between predictor variables and *Approved Conversion*. The three quantitative variables (*Clicks*, *Impressions* and *Spent*) displayed positive correlations to the response variable. For the categorical variable Age, the age between 30-34 tends to have a higher number of *Approved Conversions*. Similar trends were found with other predictor variables.

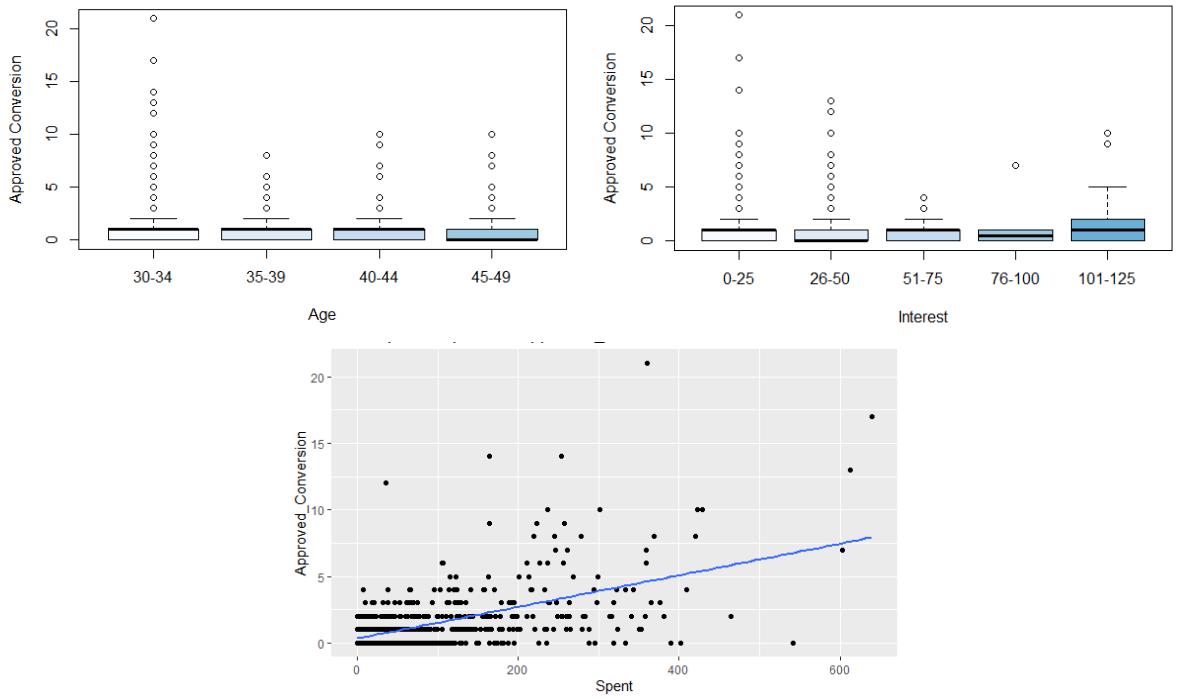

Figure 2 – Relationships between Predictor Variables and Approved Conversion

To evaluate the relationship between the predictor variables and *CTR*, boxplots and scatter plots were plotted in Figure 3. It can be observed that the values of *age* and *interest* affect the values of *CTR*, as

boxplots shift in value with categories. Additionally, there appears to be a positive linear relationship between *Spent* and *CTR*. Similar trends were observed with the rest of the predictor variables.

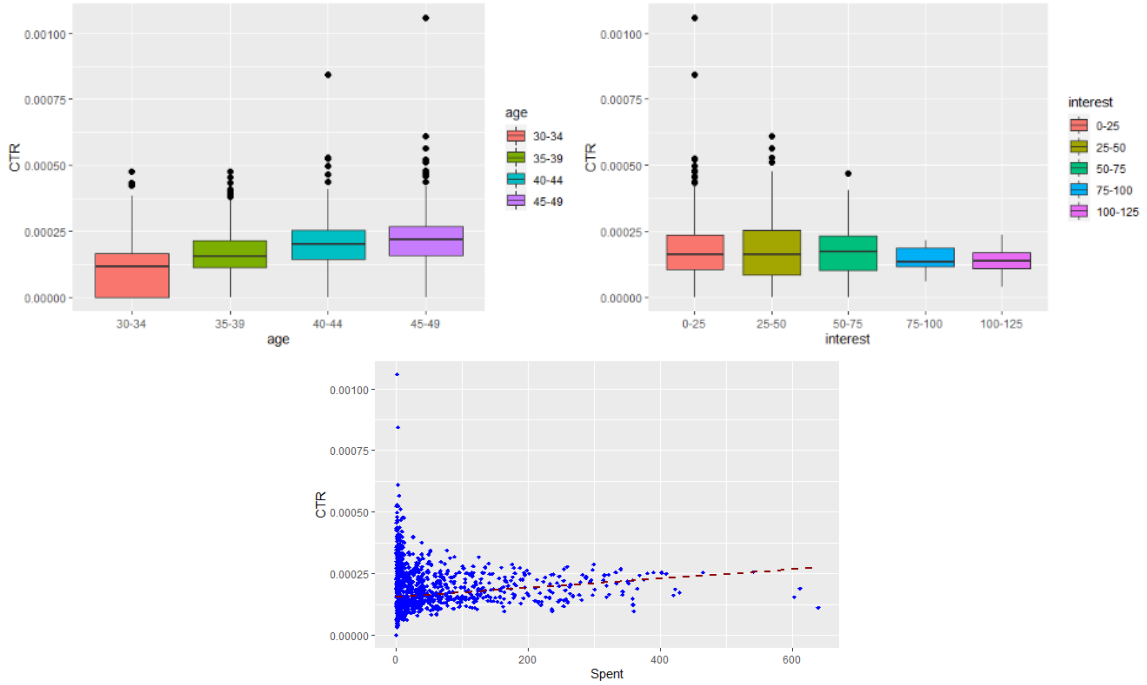

Figure 3 – Relationships between Predictor Variables and CTR

III. Methodology

As previously mentioned, in this study, we will fit two different models: (1) Multiple Linear Regression to predict *Approved Conversion*, and (2) Poisson Regression to predict *Clicks-Through Rate*. Both models were trained with 75% of the data and tested with the remaining 25%.

1. Multiple Linear Regression

1.1. Full Model

Response Variable	Predicting Variable	Type of Variable	Reference Category
Approved Conversion	xyz_campaign_id	Categorical	916
	Age	Categorical	Age 30-34
	Gender	Categorical	Female
	Interest	Categorical	0-25
	Impressions	Quantitative	-
	Clicks	Quantitative	-
	Spent	Quantitative	-
	Total Conversion	Quantitative	-

Table 3 – Multiple Linear Regression Model Summary

By the Correlation Analysis in Section II.2.2, we verified multicollinearity issues by calculating the VIF. Table 4 shows the variation inflation factor of the full model, and it further reiterates the fact that there is multicollinearity between these predicting variables with large VIF values.

	GVIF	Df	GVIF^(1/(2*Df))
xyz_campaign_id	1.526112	2	1.111467
age	1.301976	3	1.044962
gender	1.246257	1	1.116359
interest	1.359903	4	1.039174
Impressions	47.662090	1	6.903774
Clicks	117.168960	1	10.824461
Spent	221.954684	1	14.898144
Total_Conversion	3.719045	1	1.928483

[1] 10

Table 4 – Variance Inflation Factors for MLR Model

According to the p-values in the full model, *Clicks* has the most insignificant among the three variables with high VIFs: *Impressions*, *Clicks* and *Spent*. Therefore, we removed *Clicks* from the model to reduce multicollinearity issues. On a second iteration of calculations of VIFs, *Impressions* and *Spent* still displayed large VIF values. Therefore, removed *Impressions* as well from the model and solved the multicollinearity problem.

1.2. Variable Selection

To carry out variable selection and check which variables add on explanatory power one by one, we performed forward stepwise regression where the minimum model included just the intercept. The algorithm selected *Total Conversion*, *Interest*, *Clicks* and *Gender* as predictors. However, we moved forward with the full model that excluded *Spent, Clicks* and *Impressions* to do further testing. We have written about our approach and thought process in even more detail in the discussion part.

2. Poisson Regression

2.1. Full Model

Since our main objective in this is to "count" the number of clicks for a certain "spread" (here, *impressions*), we adapt the Poisson Model for this Regression Analysis by using *impressions* as the offset in the model.

Response Variable	Predicting Variable	Type of Variable	Reference Category
Clicks/Impressions (CTR)	Age	Categorical	Age Group 25-30
	Gender	Categorical	Female
	Interest	Categorical	Interest Group 0-25
	Spent	Quantitative	-

Table 5 – Poisson Regression Model Summary

The initial model results as seen in Figure 5 below indicate that all the coefficients are statistically significant except *Spent*.

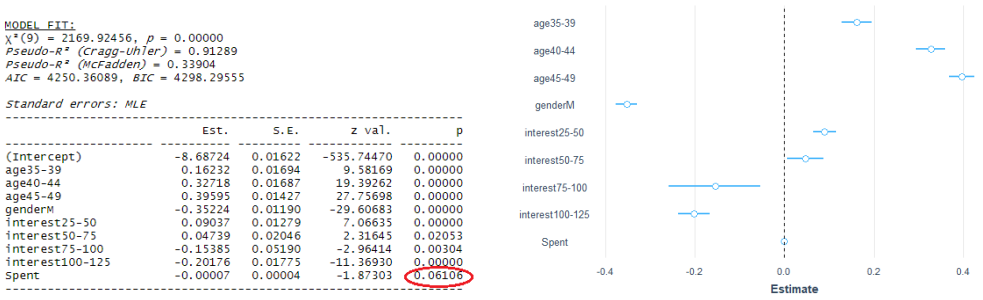

Figure 5 – R Outputs for Poisson Regression's Model

2.2. Variable Selection

First, we performed forward regression for variable selection. The model selected by this forward regression method was identical to the initial full model. It is important to note that that the variable *Spent* is not eliminated by this selection method despite it not being statistically significant. Though the coefficient value is not large (Figure 5), *Spent* it is negatively correlated to the *CTR*. Hence, if a company spends a comparatively low amount on advertising, then this investment will not have much effect on the rate of clicks. But, if the amount spent becomes large (keeping all other factors constant), the rate of clicks goes on a downside. This is probably because money is associated with the number of times a viewer is bombarded with an ad, which could irritate a viewer and he may be more likely to not click on the advertisement. We could conclude that *Spent* is one of the controlling variables in the model.

IV. Discussion of Results

1. Goodness of Fit

1.1. Multiple Linear Regression

The MLR proved to have significant goodness of fit issues. As observed in Figure 6, the variability displays an increasing trend along with the increase in fitted values. Both the full model, the model without multicollinearity issues and the model selected by the stepwise linear regression presented this issue. We moved forward with the model with no multicollinearity issues.

First, we analyzed the outliers in the dataset by calculating and plotting Cook's distance, as seen in Figure 7. The outlier in index 351 was removed. This removal did not lead to changes in the values of the model coefficients–this point was not an influential point. However, it slightly improved the overall goodness of fit.

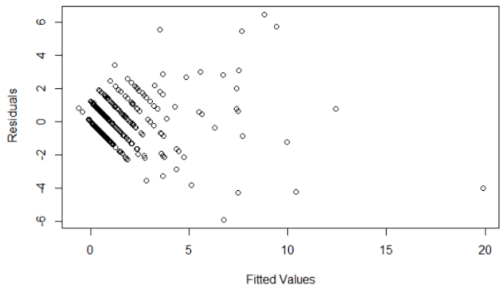

Figure 6 – Variance Assumption Plot

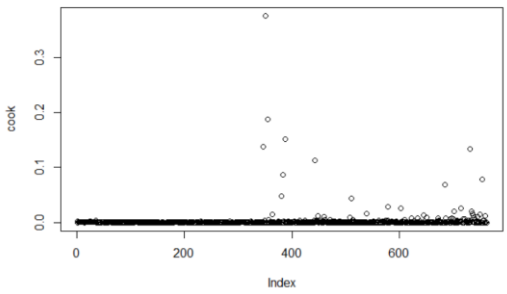

Figure 7 – Cook's Distance Plot

To further improve the fit, we applied Box Cox and multiple other transformations on the response variable. When applying Box Cox, the obtained value was $\lambda = -0.222$ which translated to an inverse transformation. However, on its application, we did not find a significant improvement: in fact, the normality assumption was even more violated. Other transformations, such as log, and square root presented similar results.

The last Linear Regression model attempted included the prediction of *Approved Conversion* with only *Total Conversion* as the predictor variable. The model's assumptions were less violated than the models mentioned previously. However, the results were still insufficient.

In short, the Multiple Linear Regression does not appear to be a suitable model to predict *Approved Conversion.*

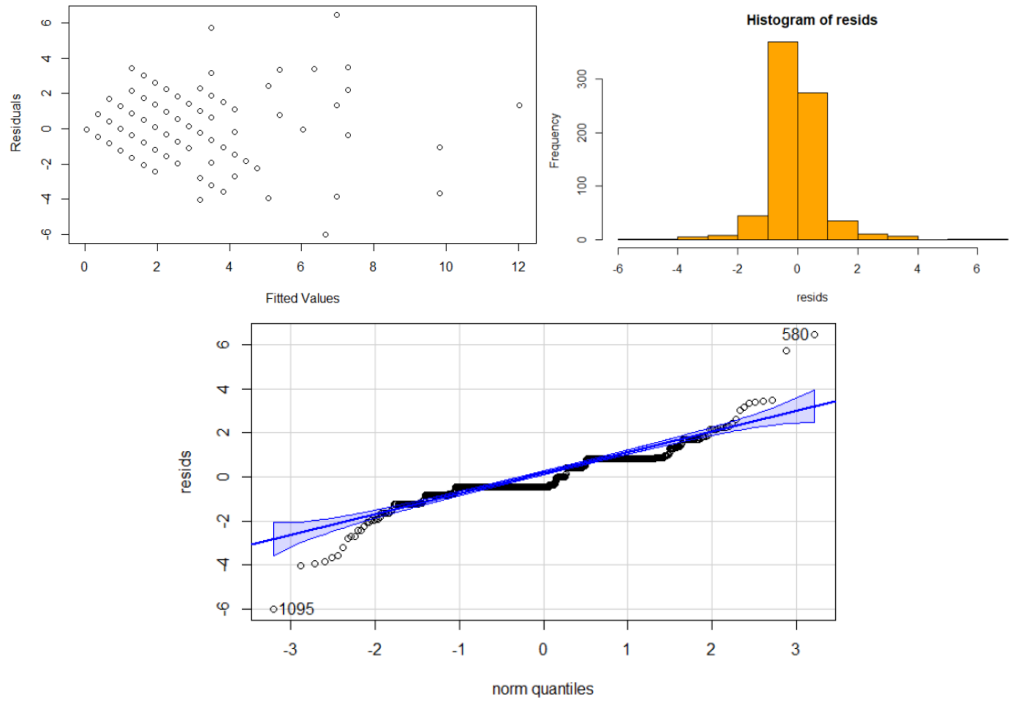

Figure 8 – Residual Analysis for MLR Model excluding Spent, Impressions and Clicks

1.2. Poisson Regression

The Poisson Regression model proved to have an excellent goodness of fit. Firstly, the normality assumption held: the Histogram of Standardized Residuals showed normally distributed residuals and the curvatures at the ends of the QQPlot were barely existent, as observed in Figures 9 (top). Second, the log of the event rate, in this case, *Clicks/Log(Impressions)*, was plotted against all predictors. In all cases, the linearity assumption held. An example of the resulting plot can be found in Figure 9 (bottom).

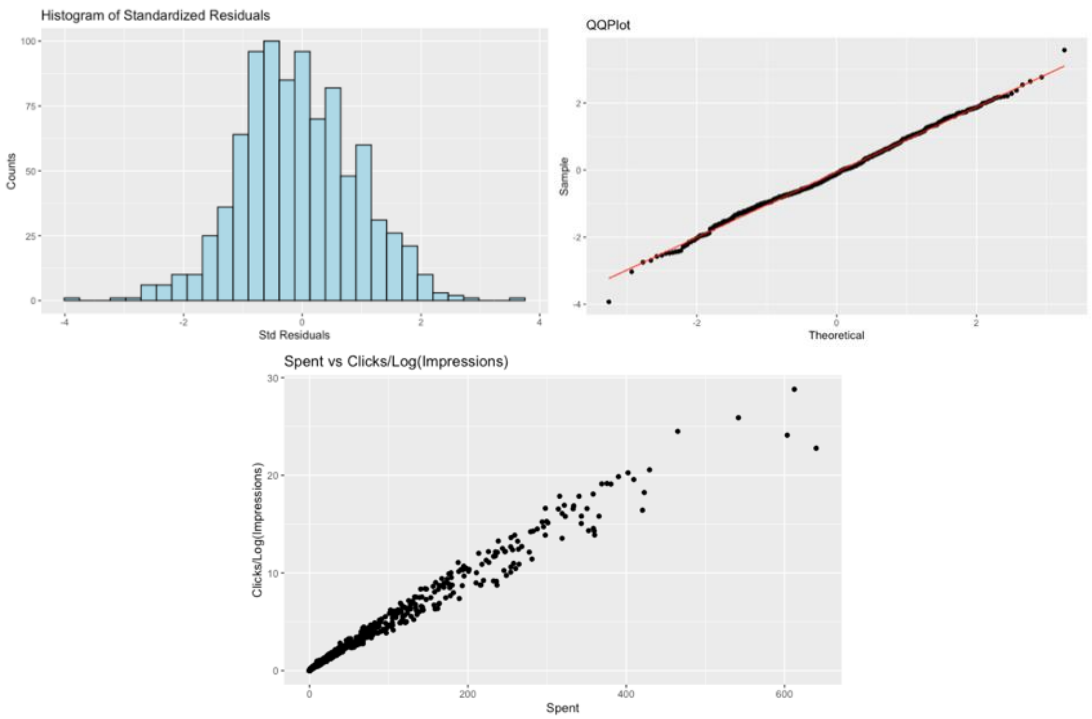

Figure 9 - Residual Analysis for Full Poisson Model

Hypothesis testing was also performed to evaluate the GOF of the model. We tested the null hypothesis H_0: *The Poisson Model fits the data* versus H_1: *The Poisson Model does not fit the data* . Using the deviance test statistic, a p-value close to 0.785 was obtained. This concludes that we do not reject the null hypothesis of good fit.

Lastly, we evaluated the model's overdispersion by calculating the overdispersion parameter. The model yielded a parameter of $\hat{\phi} = 0.962$. This verifies that the model is not an overdispersed model.

All arguments show that the Poisson Regression appears to fit the problem of predicting *CTR* very well: the model's assumptions are held, the hypothesis testing does not reject the goodness of fit, and the model is not overdispersed.

2. ***Performance Metrics***.

The Precision Measures of all models that were studied are displayed in Table 6. It can be observed that PMs for all linear models are very high, indicating a poor predictive power. The Poisson Regression exhibits the opposite pattern, the PM is close to 0.05, which validates its predictive power and accuracy at predicting *CTR*.

Model	Precision Measure (PM)
MLR Full Model	0.737
MLR Model excluding *Clicks*, *Impressions* and *Spent*	0.703
SLR Model with *Total Conversion* as single predictor	0.756
MLR Model Stepwise Regression	0.703
MLR with Inverse BoxCox Transformation	0.706
Poisson Regression	0.056

Table 6 – Performance Metrics for Tested Models

V. **Conclusion**

After testing two different models, we reached the conclusion that the relationship between conversions and other variables related to the user is not linear. The Multiple Linear Regression proved to be a poor fit to the problem despite of the multiple iterations of combinations of variables and transformations. One of the main reasons why the model underperformed could be due to the necessity of additional explanatory variables. Once a user clicks on the ad, the probability of a prospect going ahead and buying that product depends on a lot of other factors like disconnect between landing page and ad copy, uncompetitive features or pricing, usage of wrong keywords etc.

However, the Poisson Regression model successfully predicted the *Clicks-Through Rates* of for about 95% of the ads. Additionally, this model proved to be an excellent fit for this prediction because it held all model assumptions. This means that the assumed predicting variables are adequate to model the behavior of *CTR*, and that Facebook could use a Poisson Regression Model to predict the outcomes of ad campaigns to further benefit Facebook's customers and its advertisement system.

References

1. Gokagglers (2017) Sales Conversion Optimization: How to Cluster Customer data for campaign marketing. Data Repository: https://www.kaggle.com/loveall/clicks-conversion-tracking. Accessed on October 2nd 2021.