# Finding Meaningful Features to Increase Notification Engagement in a Logistic Regression Model

A Technical Report submitted to the Department of Engineering

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

**Kiran Manicka**

Spring, 2023

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Rosanne Vrugtman, Department of Engineering

# Finding Meaningful Features to Increase Notification Engagement in a Logistic Regression Model

CS4991 Capstone Report, 2023

Kiran Manicka
Computer Science
The University of Virginia
School of Engineering and Applied Science
Charlottesville, Virginia USA
km9eg@virginia.edu

**ABSTRACT**

Driving Engagement is the focus of any product focused company. For Facebook, this meant creating notifications people were likely to click on. Increasing engagement with Facebook. I developed a Machine Learning Model to create targeted notifications for users. The specific type of notification I aimed to create was a pop up on the screen that suggested a recent post created by another user. The most challenging aspect was finding features that correlated highly to whether a user would click on a notification. My primary findings included a variety of ways to measure the importance of features within a given model. This methodology can be helpful to any future machine learning work that hinges on weight the importance of various features, however this method has yet to be tested on different types of models like decision trees. This could be an important area of further testing.

## 1. INTRODUCTION

Predicting whether a notification will be clicked on or not can be summarized as being a binary classification problem. Within this context, the notification will be classified as either likely to be clicked or not likely. Binary classification offers multiple methods that can be used for this purpose. One of these methods is called Logistic Regression which centers around the idea of learning the probability distribution of the data and then predicting the target probability of a new piece of data. Generally speaking, this a very efficient method of binary classification, however it can still be useless if it is not given useful features to work with. Feature engineering revolves around finding the best and most meaningful features to train a model with. It is important for any serious data or machine learning engineer to be equipped with the proper skills of analyzing which features to incorporate into a model. Some of the methods detailed later will equip the reader with the proper toolbox for evaluating various features stacked up against each other.

## 2. RELATED WORKS

The fundamental principles of Logistic Regression were first created by Joseph Berkson in 1944 (Brett, 2020). Understanding his original paper is key to understanding completely how logistic regression is meant to be used. I chose logistic regression for this project because it is a sophisticated and

powerful yet simple method of aggregating structured data that can be used to determine a binary classification. This binary classification is choosing the label of whether a user will click on the notification or not. Often times, in binary classification task, a machine learning model won't just return a label, but a probability. A low probability would correspond to a user not being likely to engage with the notification, while a high probability infers a high likelihood of engagement. Once a reader understands the basic components of logistic regression, they are well prepared for learning about feature engineering and data analysis techniques. Data analysis methods have been around for centuries and have constantly been improving with time as computing has advanced. Studying the data analysis literature in order of chronology is recommended as it will give the reader a good sense of the development field and will also prepare the reader for this paper.

## 3. PROJECT DESIGN

First, different data manipulation techniques must be identified. Then they should all be used in separate models to measure respective performances. Lastly, the method with the best performance can be used.

### 3.1 Identifying Methods of Feature Engineering

I utilized Principal Component Analysis (PCA), Information Gain, and the Correlation Coefficient to determine the importance of each feature in the logistic regression training dataset. PCA is a method of transforming the dimensionality of the original dataset to something a lot smaller. The idea is that the new and smaller dimensionality will be easier to compute and will still contain the importance and variance that the original dataset had while also being easier for the model to understand. Information Gain is used a lot with decision tree models. This method looks at each feature and calculates how much it explains the output feature. A higher information gain correlates with the output feature being more dependent on the input feature. The Correlation Coefficient is a method that is similar to information Gain and measures how much a feature is connected to the output.

### 3.2 Different Methods

Once a method is selected, there are two different ways to incorporate this into the model: Feature Weighing and Omitting unimportant Features.

#### 3.2.1 Feature Weighing

The first method involves changing the weighting of the features based on the perceived importance. This means that more important features will impact the model more during training while less important feature will have the opposite effect. Less important features will still impact the model even though it might not be by that much.

#### 3.2.2 Omitting Unimportant Features

The second method involves just leaving out the unimportant features completely. The hope is that this will make training more computationally efficient while also letting the important features decide the output.

### 3.3 Evaluating Performance

The last step in this procedure is to measure the performance on data the model has never seen before. There were 3 methods outlined earlier for measuring feature importance as well as 2 methods to implement them. The process that I used to compare these methods was to use somewhat of a grid search

technique. This essentially means trying out all the possible combinations of techniques and then picking the best performing model. This makes the process easer since it is objective instead of trial and error. Since there were 3 methods of finding feature importance, and 2 methods of implementing these features, there were 6 possible combinations. The combination that yielded the highest performance for the Logistic Regression model was using the Correlation Coefficient method with Omitting Unimportant Features.

## 4. RESULTS

Utilizing the Correlation Coefficient method as well as omitting unimportant features helped the Logistic Regression model go from an accuracy of 63% to 75%. The result was interesting not in the sense that the accuracy went up, but in the fact that the Correlation Coefficient and Omitting Unimportant Features ended up being the most efficient combination. The reason this is peculiar is that these methods are very simple. I believe that the simplicity of these methods helped the model perform well. In a lot of cases overcomplicating model architectures and data processing steps can end up confusing the training process.

## 5. CONCLUSION

The improved performance of the model shows that some features are not necessarily essential to the model learning the distribution of the data. By slicing these features out or performing some sort of feature engineering to create a new meaningful input, the machine learning model can learn a lot more and improve on accuracy or it can become more computationally efficient. I hope that this paper can serve as an aid to people interested in data science or current machine learning engineers who are interested in boosting the performance of their models. Overall, I believe this methodology is extremely important to large and complex models that need every extra bit of accuracy they can get.

## 6. FUTURE WORK

In order to expand this project and this research, I would like to not only dive deeper into the technical aspects of feature engineering, but also study and observe the effects of hyper parameter tuning. There are many different types of hyper parameters that can alter the performance of the model, and these, in conjunction with feature engineering, give even greater control the machine learning engineer.

## REFERENCES

Brett, D. (2020). Exploratory Data Analysis of NYPD Arrest Data [RStudio Cloud Project]. Retrieved May 2, 2023, from https://rstudio-pubs-static.s3.amazonaws.com/856695_1a27bf9a92ed489eb5520e181dc6b204.html

German, A., & Hill, J. (2007). Data analysis using regression and multilevel/hierarchical models. Cambridge University Press.

Pandey, R. (2021). A Comprehensive Guide to Feature Engineering for Machine Learning. Analytics Vidhya. Retrieved from https://www.analyticsvidhya.com/blog/2021/06/a-comprehensive-guide-to-feature-engineering-for-machine-learning/