Data Mining
Final Project

Classifying User Intent from Click Data

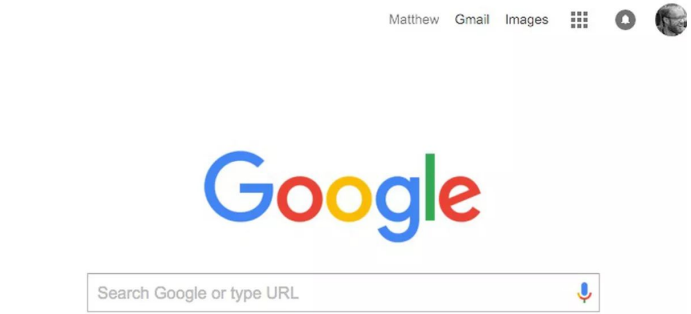
Arianna Lupi

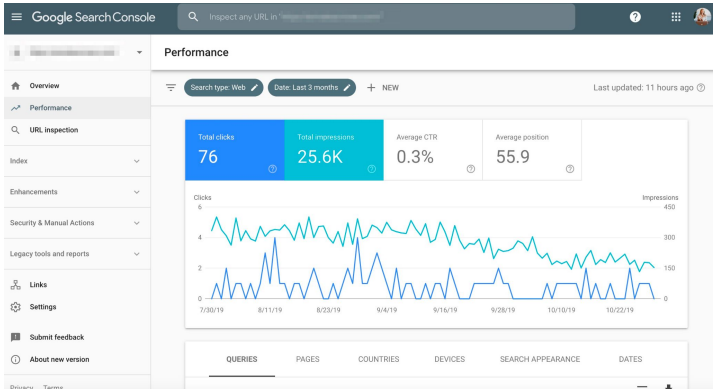
Context

As a search marketer, I find myself looking every day at different data sources that describe the user journey across a website.

Google provides anonymous data on user behavior, from the keyword they used to find a specific page, to how long they stayed in it, to which page they jumped next etc.

Sometimes, it's hard to take decisions and forecast growth.

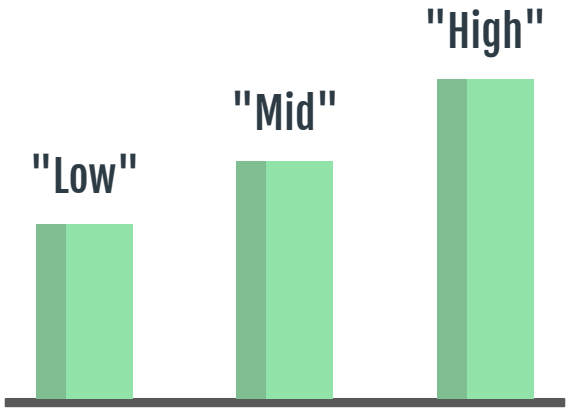




The problem

Can we predict the keyword intent from a user's click data?

Intent for the purpose of this research will be classified in



according to SEO industry standard data.

The Data

This data is a csv exported from an seo agency's Google Organic Search data. There's 1,000 rows of information on specific keywords, categories, clicks, impressions, click through rates, and conversions.

We can see this data set has the following columns:

Keywords — The keywords or key phrases users found to get to the agency's website.

Category — The category of the topic of the keyword, if it's either about SEO, SaaS, Marketing, Link Building, if it's a Landing Page or other if it falls in another category.

CTR — % of users that clicked through the page after an impression.

Clicks — Clicks made to the page form the given keyword.

Impressions — Impressions to the page from a given keyword.

Position — The position of the page for a given keyword in Google's result pages.

Conversions — Number of conversions for a given keyword

Data Source: Google Data Studio

Data Transformation

We transformed the data by adding a new column called “Intent”

Intent is classified the following way:

High Intent

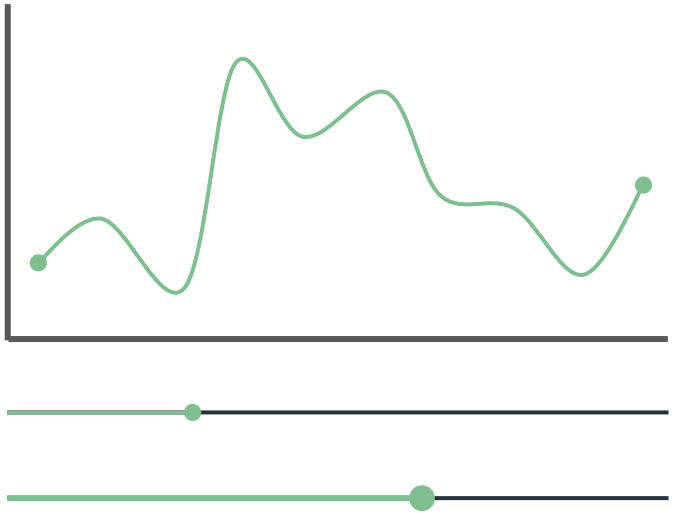
keywords that include the terms “best|agency|tools|consultant|consultancy” Users are looking to buy and make a purchase.

Mid Intent

keywords that include the terms "how|what|who|when|where|why|which” users have a navigational intent and are looking to solve a problem or get answers to a question.

Low Intent

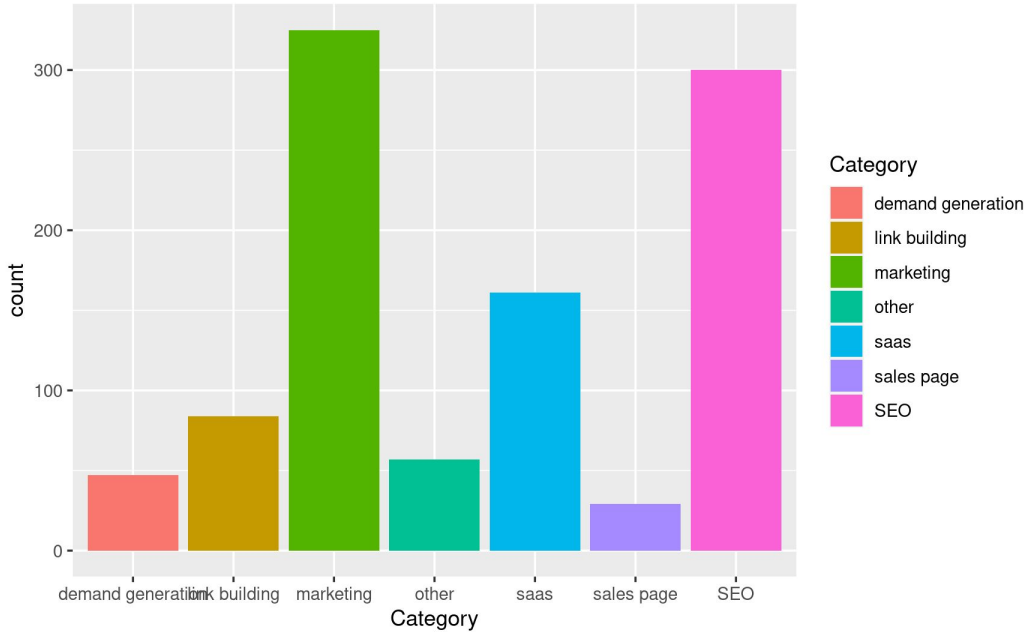
All the others not classified in High or Mid



Data Visualizations

Data visualization

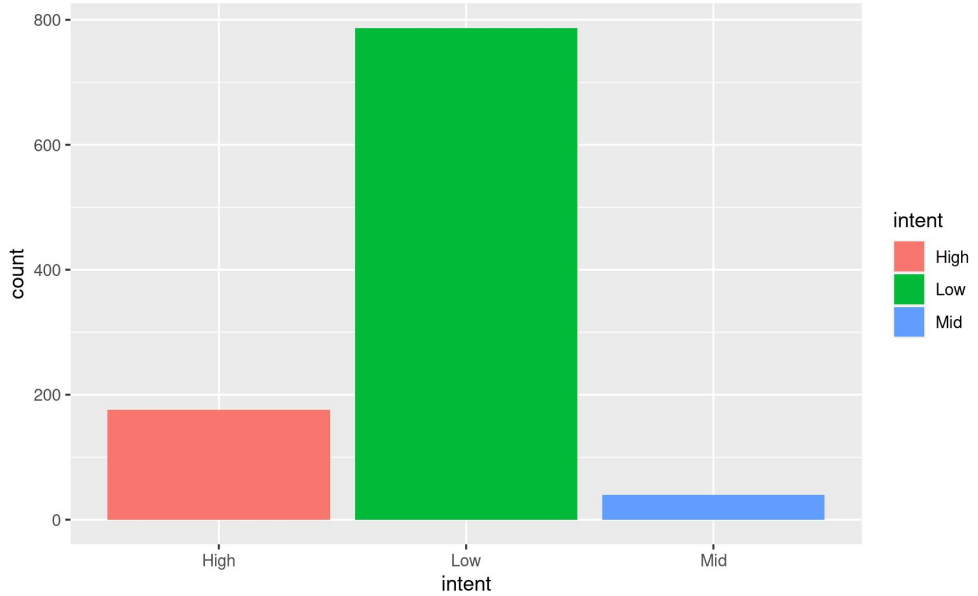
Count of Keywords per Category



We can see here in this visualization that the Category with the most keywords is Marketing

Data visualization

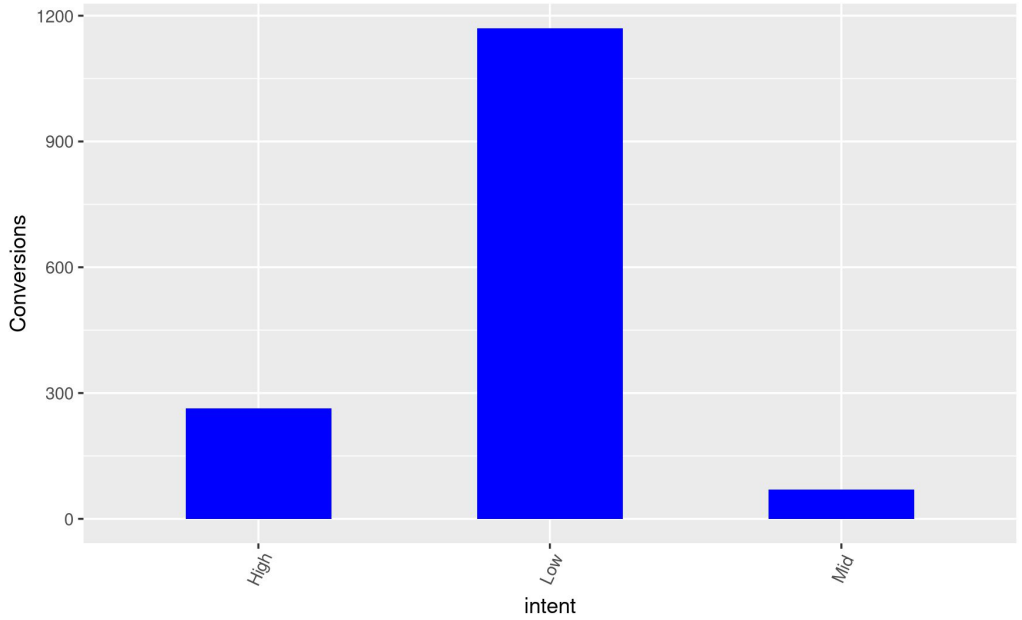
Count for each type of Keyword Intent



We can see form this bar plot that the intent with the most keywords is "Low"

Data visualization

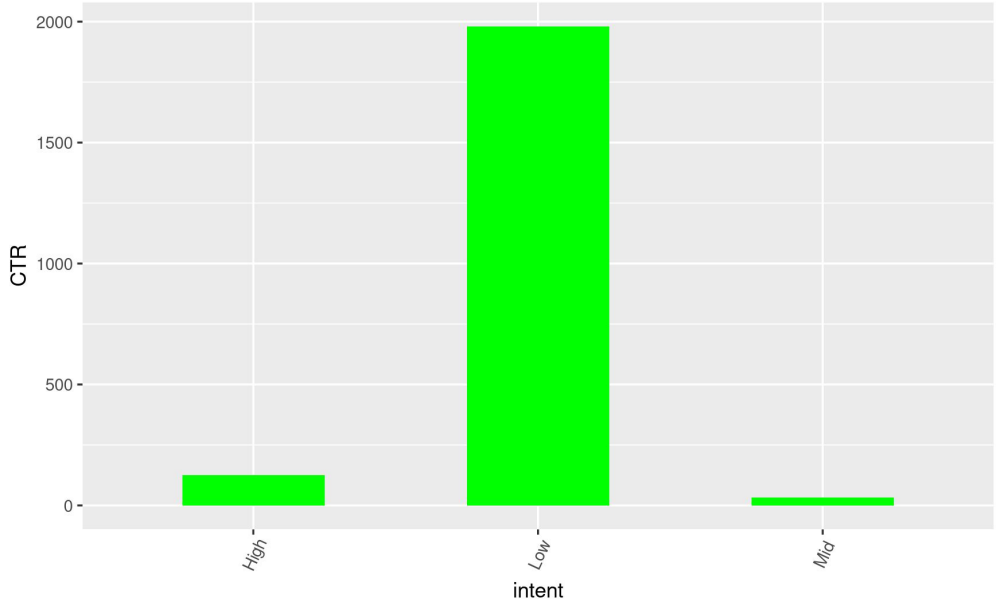
Conversions per Intent



Since there are more “Low” intent keywords there are more low intent conversions.

Data visualization

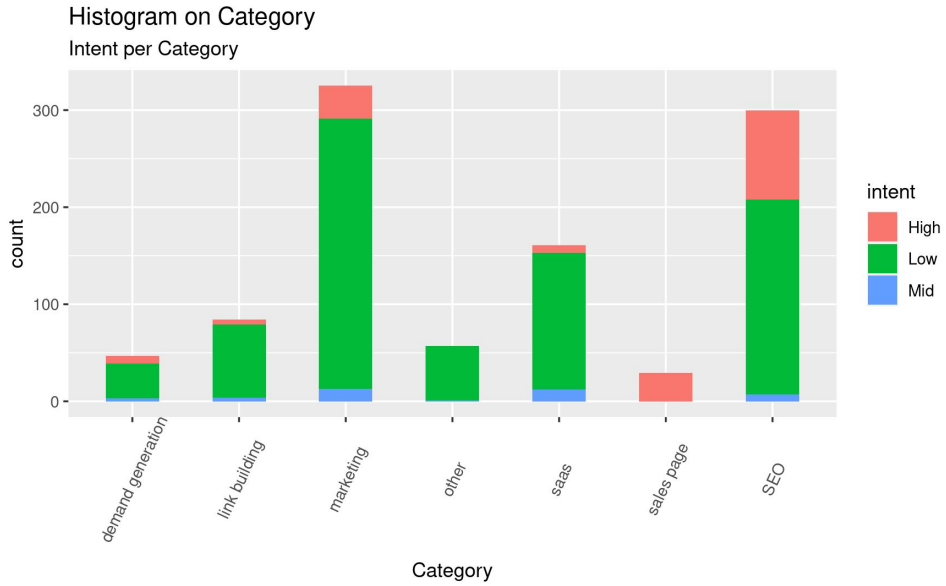
CTR for each Intent



The sum of all low intent keywords have the biggest CTR

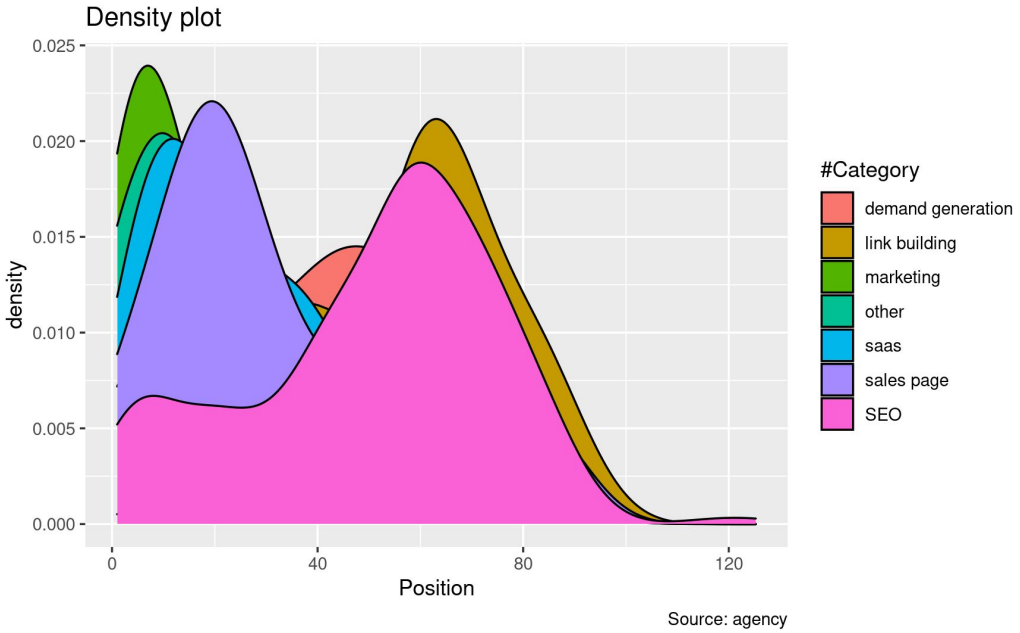
Data visualization

Intent per Category



The keyword category with the higher intent is SEO

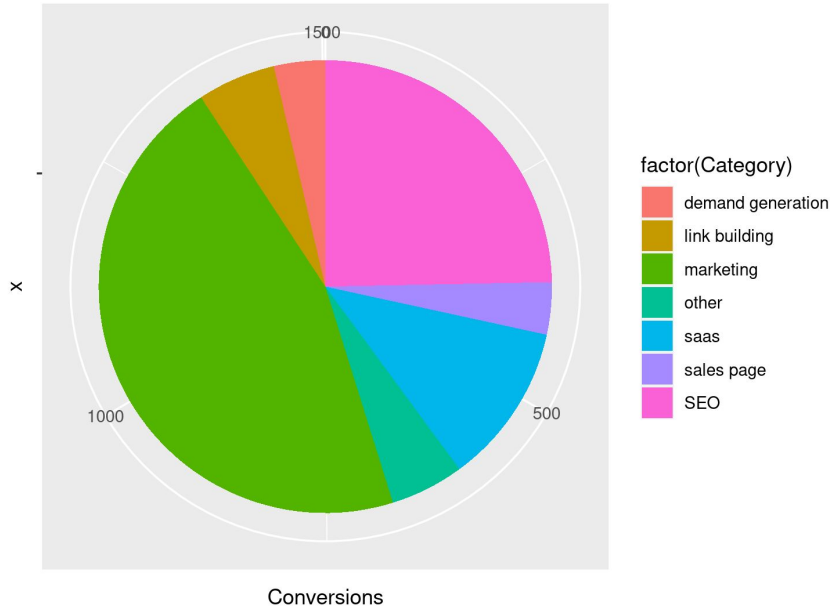
Data visualization



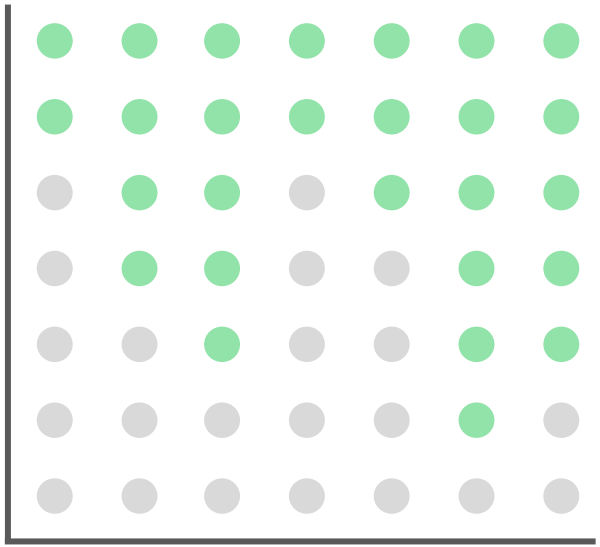
We can see here most Marketing keywords rank between 0-20
SEO Keywords Rank from 0 to 120

Data visualization

Category and Conversions Pie Chart



The Marketing Category takes most of the conversions.



Methods Used

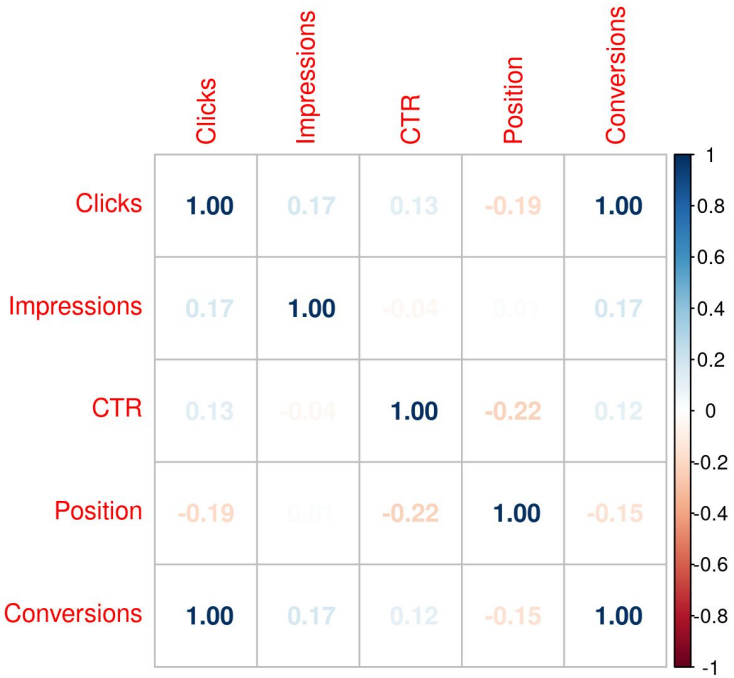
Linear Regression & StepWise Model

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable

Stepwise regression is the step-by-step iterative construction of a regression model that involves the selection of independent variables to be used in a final model. For this model, the stepwise regression selected all of the variables.

- Linear Regression RMSE: 6.311982
- Stepwise Backward RMSE: 6.311982
- Stepwise Forward RMSE: 6.311982
- Stepwise Both RMSE: 6.311982

Correlation Matrix



From the Correlation Check, I'm choosing Clicks and Conversions since they are highly and positively correlated

A correlation matrix is simply a table which displays the correlation coefficients for different variables. The matrix depicts the correlation between all the possible pairs of values in a table.

KNN

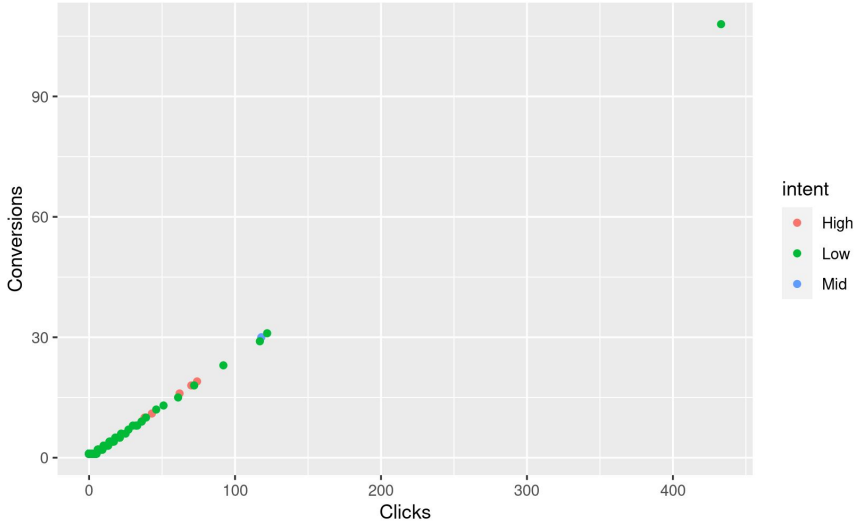
A K-Nearest-Neighbor algorithm, often abbreviated KNN, is an approach to data classification that estimates how likely a data point is to be a member of one group or the other depending on what group the data points nearest to it are in.

From the variables chosen by the on the regression method, I will create a K Nearest Neighbors model using the variables with the highest correlation and display the accuracy

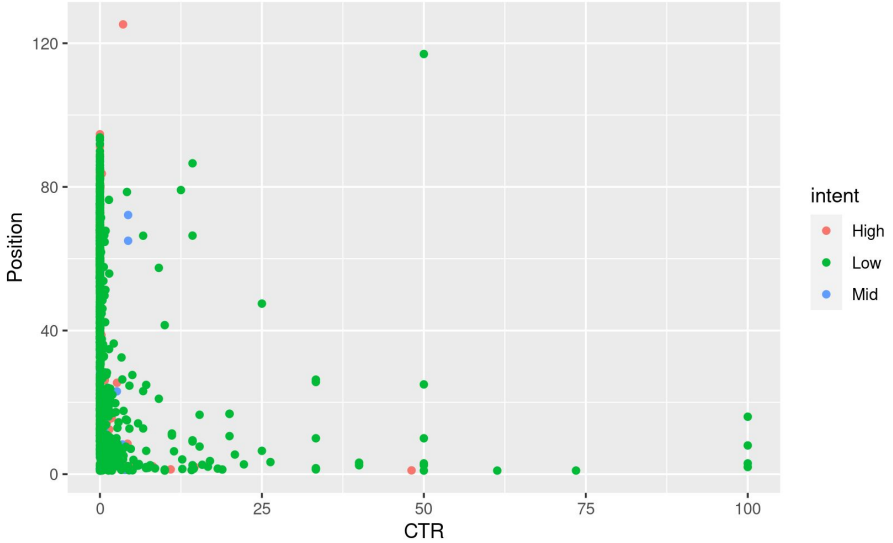
Although from part one, all Stepwise models output the same RMSE, used Stepwise Forward as it contains the least amount of variables.

KNN

Conversions and Clicks

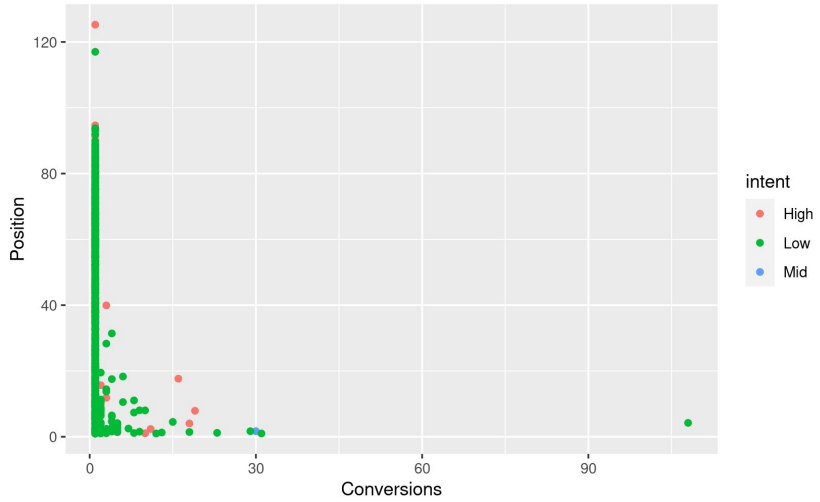


Position and CTR

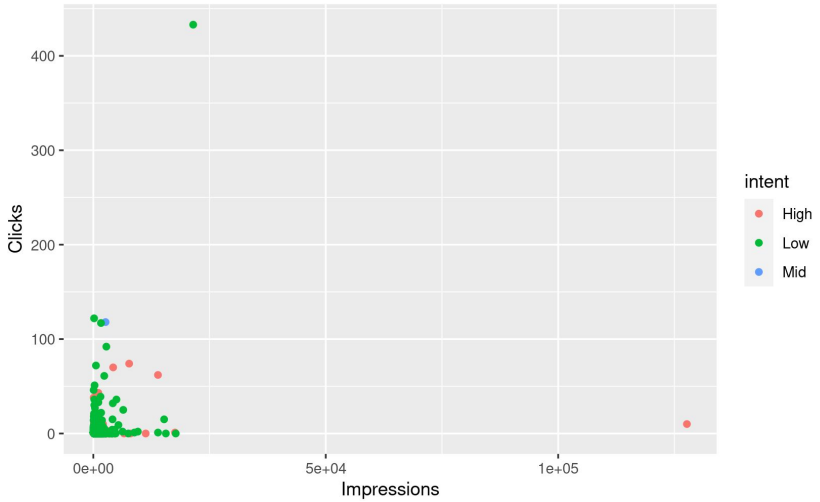


KNN

Conversions and Position

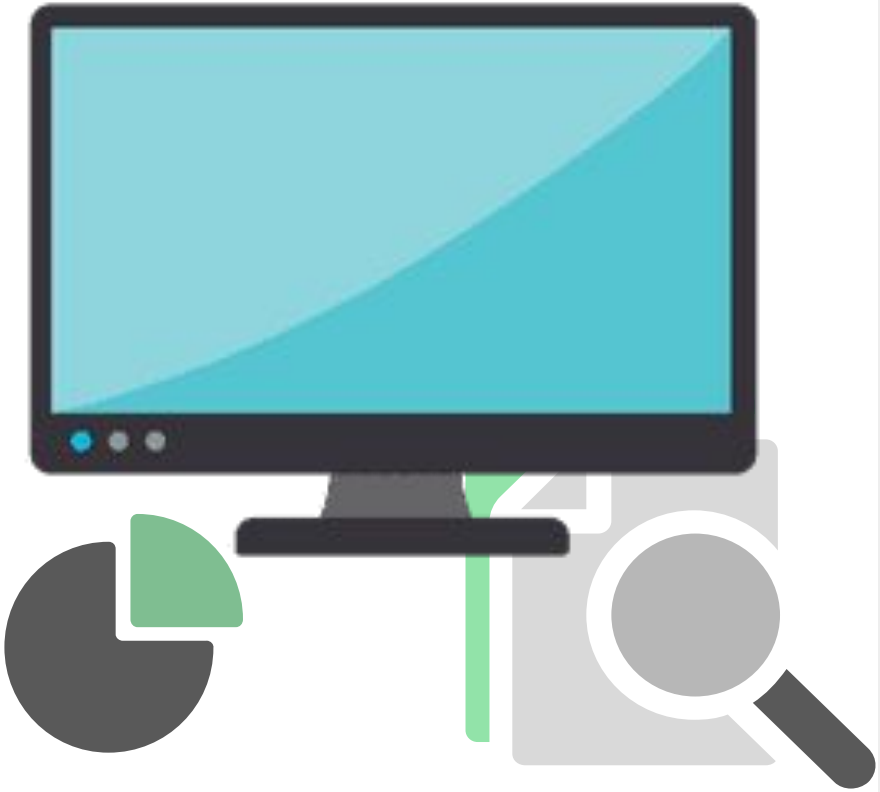


Clicks and Impressions



KNN Accuracy

Variable1 <chr>	Variable2 <chr>	accurracy <dbl>
Clicks 4 rows	Conversions	0.7830424
CTR	Position	0.7356608
Conversions	Position	0.7605985
Impressions	Clicks	0.7157107



Thank you