**Project Attribution Guideline - (Last click attribution)**

1. Goal:

* find attribution credit to assign to each veritone related in-driect orders, find how much percentage to take from each transaction
* Evaluate different goals of a customer separately (e.g. zola: invitations/ registry\_sign\_ups/ wedding\_website\_sign\_ups)

1. Method:

* Data plot: plot time decay curve

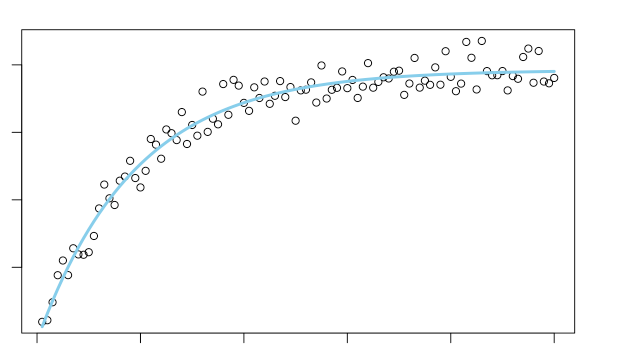
X : time difference between impression and conversion

Y: percentage non-converted

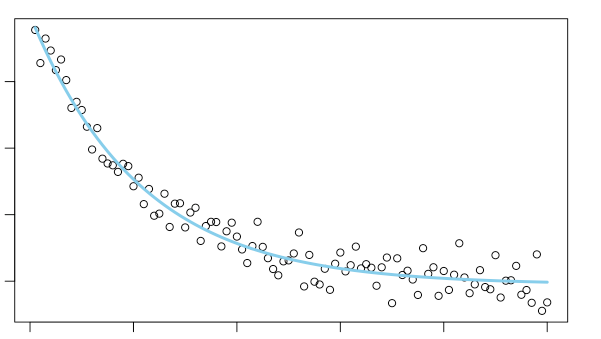
* Build a regression model to fit the training dataset, use percentage non-converted as target variable, and time difference as predictor variable, also take other variables such as placement level, channel type, genre, device etc into consideration

1. Model Evaluation & Feature Selection

* Divide the dataset into training and test group
* Perform feature selection using L1 regularization to define variable importance
* Exclude non-important variable and check r-square
* Run model on the test set to check model efficiency



Y ~ Alpha \* Exp(Βeta1X1+ Βeta2X2+ Βeta3X3 +…+ ΒetaNXN + Theta)



Time difference

**Molina - Pandora**

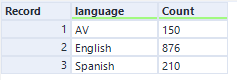
1. Sample size: 1236 samples
2. Variables

* Predicted Variable

1. Month ( 11,12)
2. Type: Audio, Display
3. Device: Android, iOS, AV
4. Language: English, Spanish
5. AudienceL Below 50, Display, Latin Genre
6. State: California, Mississippi, South Carolina

* Target Variable: difference\_in\_sceonds

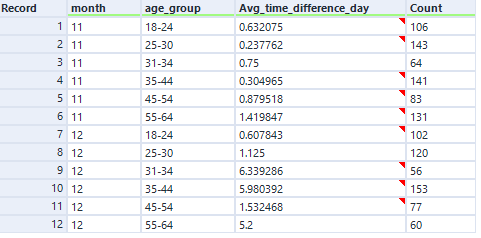
**Data Exploration:**



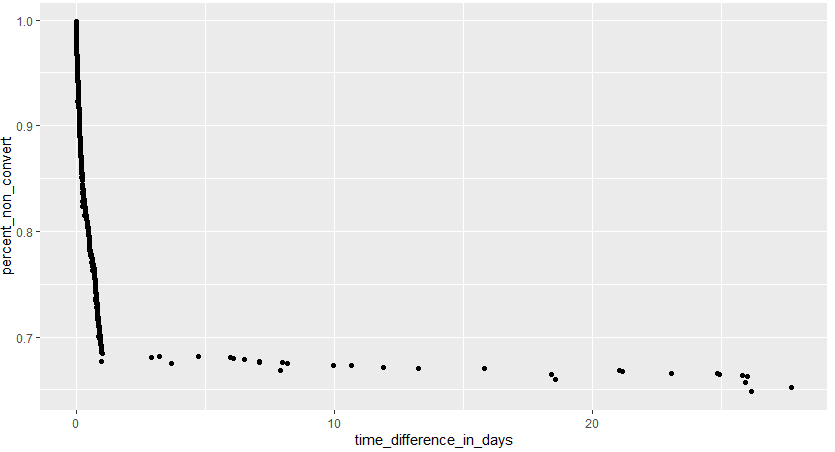
* Most people listen to english channel



* Lead conversion is much faster in Nov compared to Dec

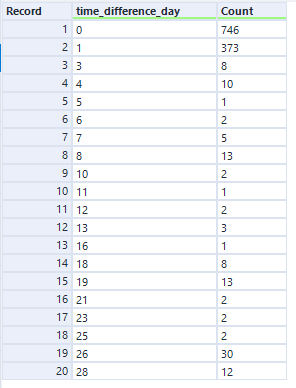


* Younger age groups are faster in conversion than older group



* Most conversions happen within the first couple of days, considering the tail spot as outliers that might impact model efficiency, we would remove outliers.

Outliers (meaning more likely to be a unique case)

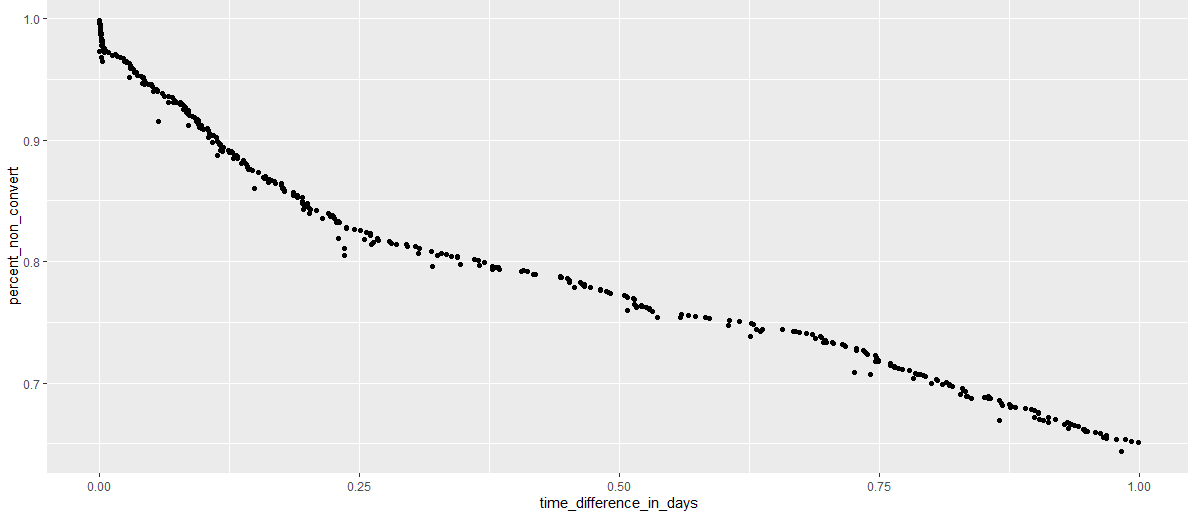


Most of the conversions are happened within day 1:



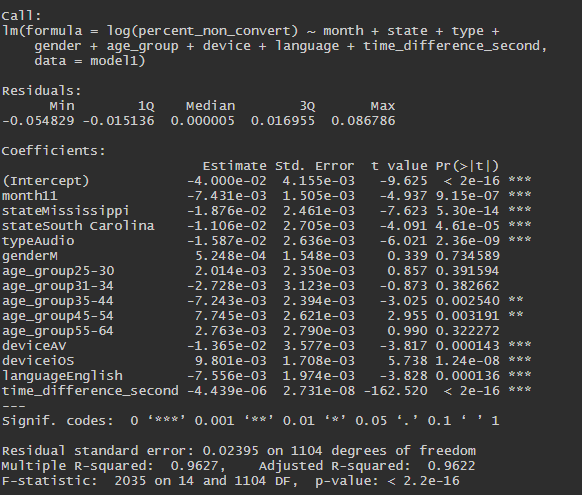
We will filter out all those converted after day 1, which we captured 91% of the whole population.

**After removing outliers:**



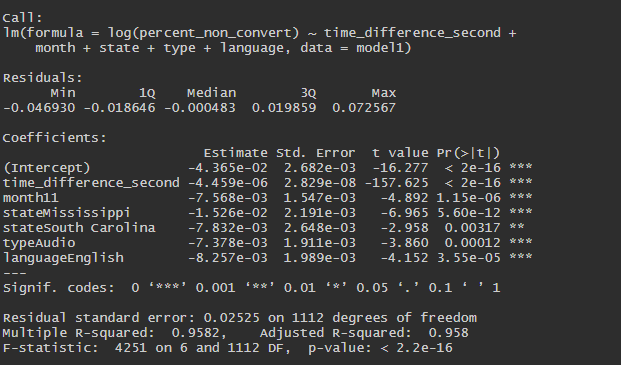
**Run Model:**

With all potential variables plugged in:

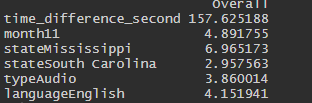


For age\_group, we tried a few different new groupings, but can’t make all levels of the age\_group statistically significant. Keeping only a few levels and removing others would be problematic, the variable has to either be considering or not considering as a whole.

Gender and device are not statistically significant as well. So we removed age\_group, gender and device, and reran the model:

New model with R-square of 0.96

**Variable importance:**



Ran variable importance test, the number explained that time is the biggest influencer on attribution rate.

**Formula:**

* To track how much percentage of credit to take for each order, plug in the coefficient to the formula to get the final result.
* Base line is month = 11, State = “California”, type = “Audio” and language = “English”
* **exp((-4.365e-02) + time\_difference\_second \* (-4.459e-06) + month11 \* (-7.568e-03) + stateMississippi \* (-1.526e-02) + stateSouth Carolina \* (-7.832e-03) + typeAudio\*(-7.378e-03) + languageSpanish \* (-8.257e-03)) \* 1.044**

**Molina - Audiology**

1. Sample size: 1887 samples
2. Variables

* Predicted Variable

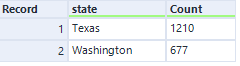
1. Month ( 11,12)
2. Device: Android, iOS, AV
3. Language: English, Spanish
4. Audience: Below 40, In Market
5. State: Texas, Washington

* Target Variable: difference\_in\_sceonds
* Pandora

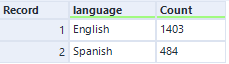
**Data Exploration:**



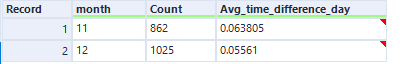
* Audiology only provides sample that converted within 1 days, within most converted in day 1



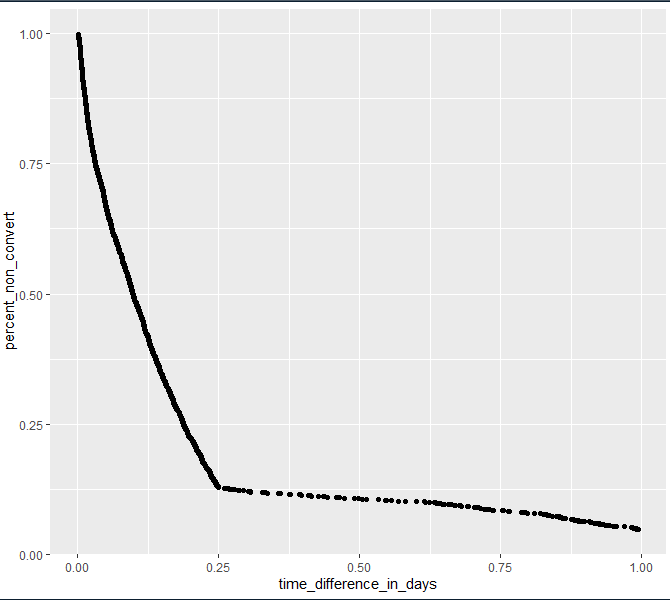
* 2x the audience size from Texas than from Washington



* Most people listen to english channel



* More lead conversions in December, and conversion speed is faster than in November



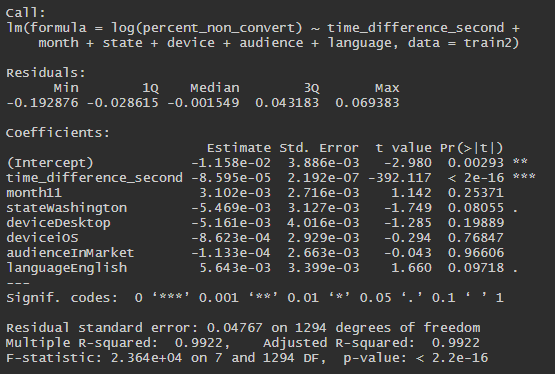
* Scatter plot illustrates that before day 0.25 (6 hours), it’a a convex curve in time decay, but after day 0.25, it’s more like a linear relationship between predictive variables and target variable. The final model would be two parts, with before day 0.25 as a time decay, and after day 0.25 a linear regression



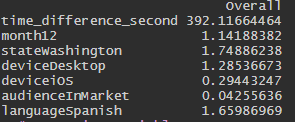
**Run Model:**

**First part: Within 0.25 day**

With all potential variables plugged in:

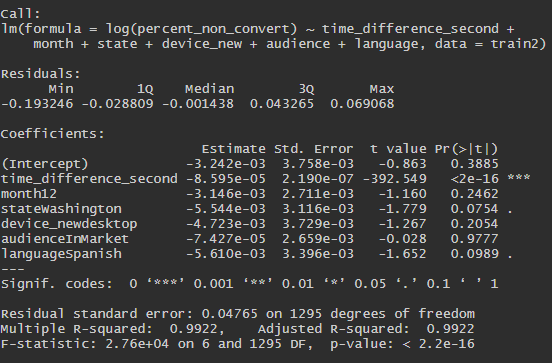


**Model Importance**

****

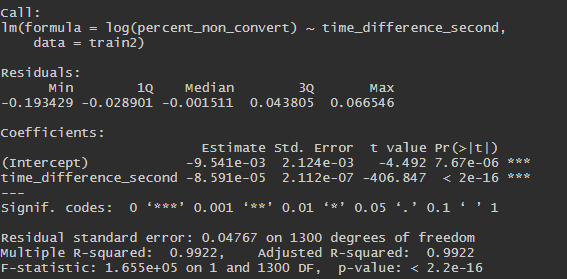
**Regrouping**

For device, we tried a few different new groupings, grouped “Android” and “iOS” under “mobile”, and others under “desktop”



New model with R-square of 0.99. Time plays the most important role, and other variables not significant on a significant level of 0.05

**Run model on time\_difference\_second only:**



**Apply model on test data:**

Correlation Accuracy:

1. Model 1: 98.17%
2. Model 2: 98.15%
3. Model 3: 98.17%

* Conclusion: time\_difference\_min plays the most important role, removing other variables didn’t lower down the accuracy rates.

**Formula:**

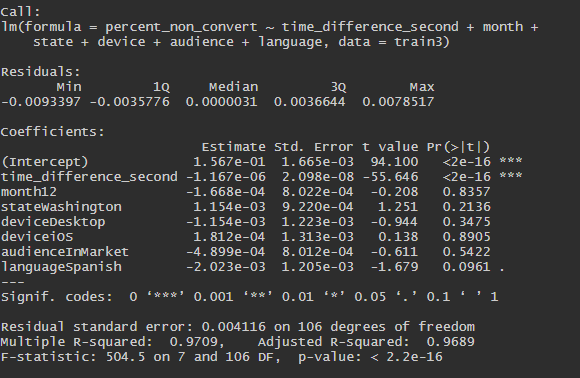
* To track how much percentage of credit to take for each order, plug in the coefficient to the formula to get the final result.
* **y= exp((-9.541e-03) + time\_difference\_second \* (-8.591e-05))**

**Run Model:**

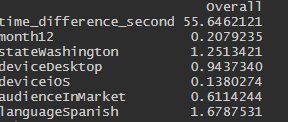
**Second part: After 0.25 day(Linear relationship)**

With all potential variables plugged in:

Model 2.1:



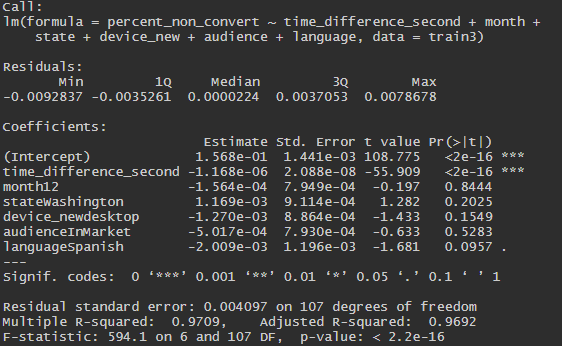
**Model Importance**

****

**Regrouping**

For device, we tried a few different new groupings, grouped “Android” and “iOS” under “mobile”, and others under “desktop”

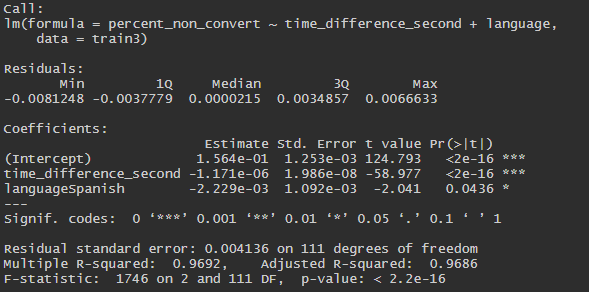
Model 2.2:



New model with R-square of 0.97. Time plays the most important role, and other variables not significant on a significant level of 0.05

**Run model on time\_difference\_second only and language**

Model 2.3



**Apply model on test data:**

* Correlation Accuracy:

1. Model 2.1: 98.17%
2. Model 2.2: 98.33%
3. Model 2.3: 98.40%

* Conclusion: time\_difference\_second plays the most important role, variable language is also statistically significant. removing non-significant variables didn’t lower down the accuracy rates.

**Formula:**

* To track how much percentage of credit to take for each order, plug in the coefficient to the formula to get the final result.
* **y= (1.564e-01) + time\_difference\_second \* -1.171e-06 + language\_Spanish \* (-2.229e-03)**

**Part I +Part II**

* **y= within\_0.25\_day \* (exp((-9.541e-03) + time\_difference\_second \* (-8.591e-05)) ) + not\_within\_0.25 day ((time\_difference\_second-21600) \*( -1.171e-06) + language\_Spanish \* (-2.229e-03)))**

**Reference:**

A:\Analytics\Client Folders\Molina\Time Decay - Pixel Attribution