Predicting Online Credit Card Applications

Digital Advertising

Lee Morgan
February 7, 2017

Agenda

Predicting Online Credit Card Applications
Digital Advertising

1. Introduction
2. Dataset
3. Exploratory Data Analysis
4. Machine Learning
5. Business Metrics
6. Conclusions and Next Steps

Introduction

Problem Statement

- Increase online applications for credit cards
- Determine what features of online ads may influence users to apply for credit cards online

YouTube *Video* → Cash Card *Display 300x250* → Clicked *Display 300x250* → Remarket *160x600* → **Account Opened?**

Travel Card *Display 728x90* → Mobile Ad *Display 300x50* → Rich Media *160x600* → **Account Opened?**

Hypothesis

Users will be more likely to apply for a credit card when they are shown more ads, specifically from ads that are:

- **"Upper-Funnel"**
- **Video**
- **"Viewable"**

Dataset

Original Datasource

- **Time frame:** Nov 8-Dec 31, 2016
- **Scope:**
    - Canadian financial services client
    - Ads bought through DoubleClick Bid Manager
- **Impression / Click / Application data**
    - Event time
    - Campaign
    - Creative ID (map to type and size)
    - Device Type
    - Viewability

Row	Event_Time	User_ID	Advertiser_ID	Campaign_ID	Ad_ID	Rendering_ID
1	1482825405273106	AMsySZbgfoIsr_614S1xlTibWfUj	[redacted]			
2	1478841372281499	AMsySZb1YchOvcxAsr3OOA7Eu7Tc	[redacted]			
3	1478673893487277	AMsySZarJkbgFRHH8_gcnYZipRIE	[redacted]			
4	1479285100173070	AMsySZYVATQGkQGi254NNUWuhrfT	[redacted]			
5	1480521490733703	AMsySZZBghGWB7qkuxQZnUL4doHD	[redacted]			
6	1478827152663769	AMsySZak7mKuDHmA4bjzzCBe-kBD	[redacted]			

Initial Feature Engineering

- Create dataset at the **user-level** by summarizing their ad exposure
- Summary of new engineered features:
  - Time difference between first and last ad
  - Total impressions
  - Impressions by:
    1. Campaign strategy
    2. Creative message / card type
    3. Creative type
    4. Creative size
    5. Device type
  - Viewable impressions
  - Clicks
  - Applied for credit card

See appendix for full set of features

Row	User_ID	Impressions	TimeDiff_Minutes	TimeDiff_Minutes_AVG	Funnel_Upper_Imp	Funnel_Middle_Imp	Funnel_Lower_Imp	Campaign_Message_Travel_Imp
1	AMsySZb5URoHQAqFtc2yx7eWq2AQ	4	9	3.0	0	4	0	0
2	AMsySZZBemBdfIkICNi3QoUi495D	2	39	39.0	0	2	0	0
3	AMsySZYC0gKN-GICxK2WHC9VbmRV	4	301	100.33333333333333	0	4	0	0
4	AMsySZZYuKRxsvW7VFSOGRWIsYZ6	1	null	null	0	1	0	0
5	AMsySZarmBmNJttVh1RdvZNIN7d5	3	103	51.5	0	3	0	0
6	AMsySZZF6A8-Mo46fGpuijpIL7cP	1	null	null	1	0	0	1
7	AMsySZaOxWidhMNLX5hVPrNdHPc7	1	null	null	0	1	0	0

Exploratory Data Analysis

Predicting a very rare event

74.1M
Ads

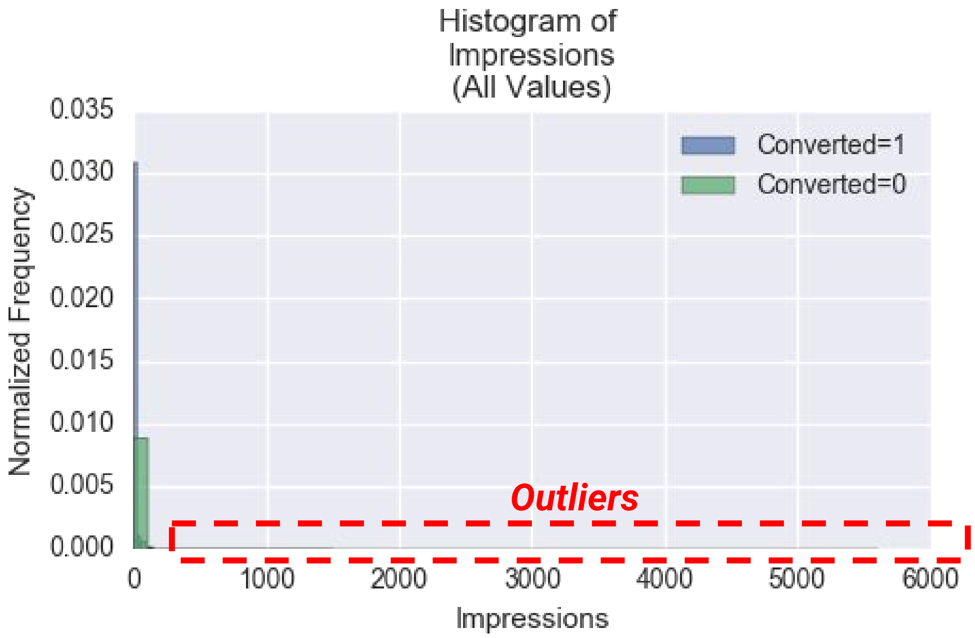
15.2M
People

1,974
Applications

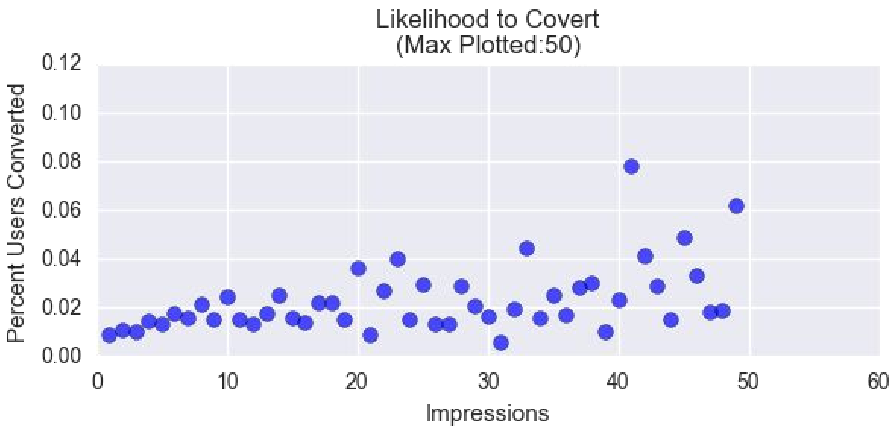
0.0013%
Conversions per User

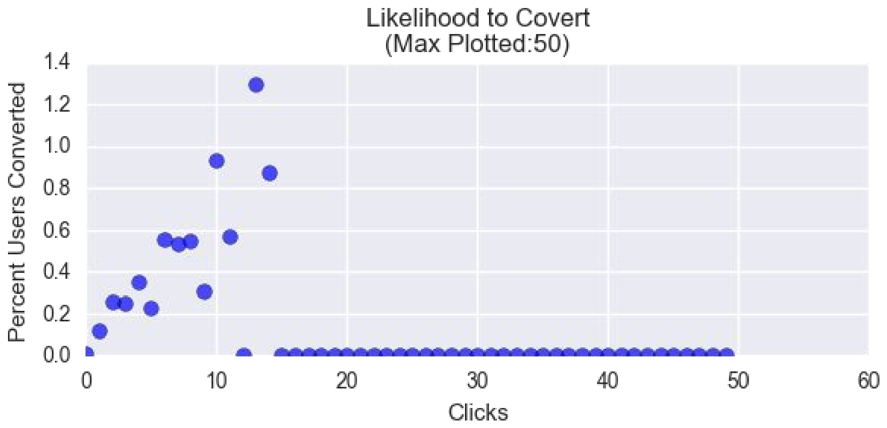
EDA - Impression Distribution

Outliers with 1000s of impressions
Removed users that received 3x standard deviation (~50 ad Impressions)

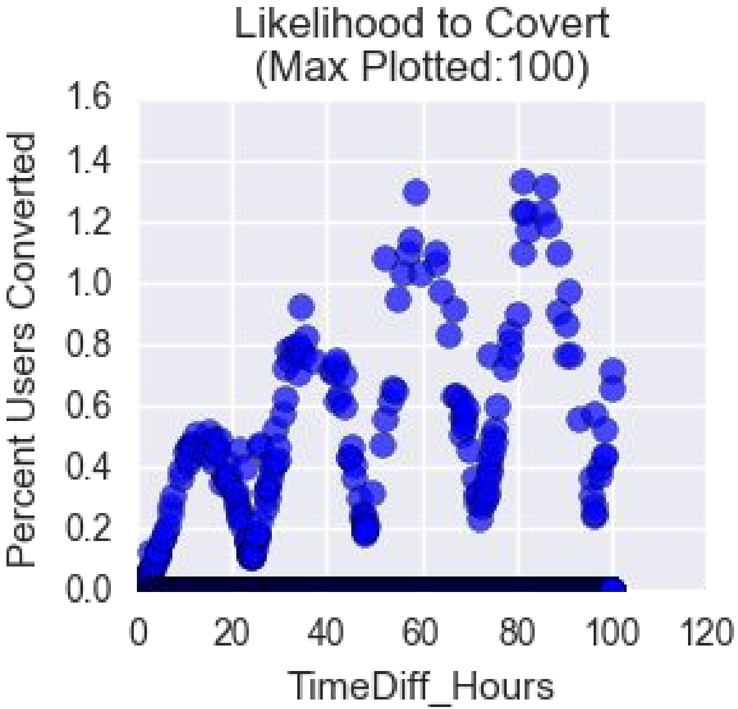


EDA - Impressions / Clicks



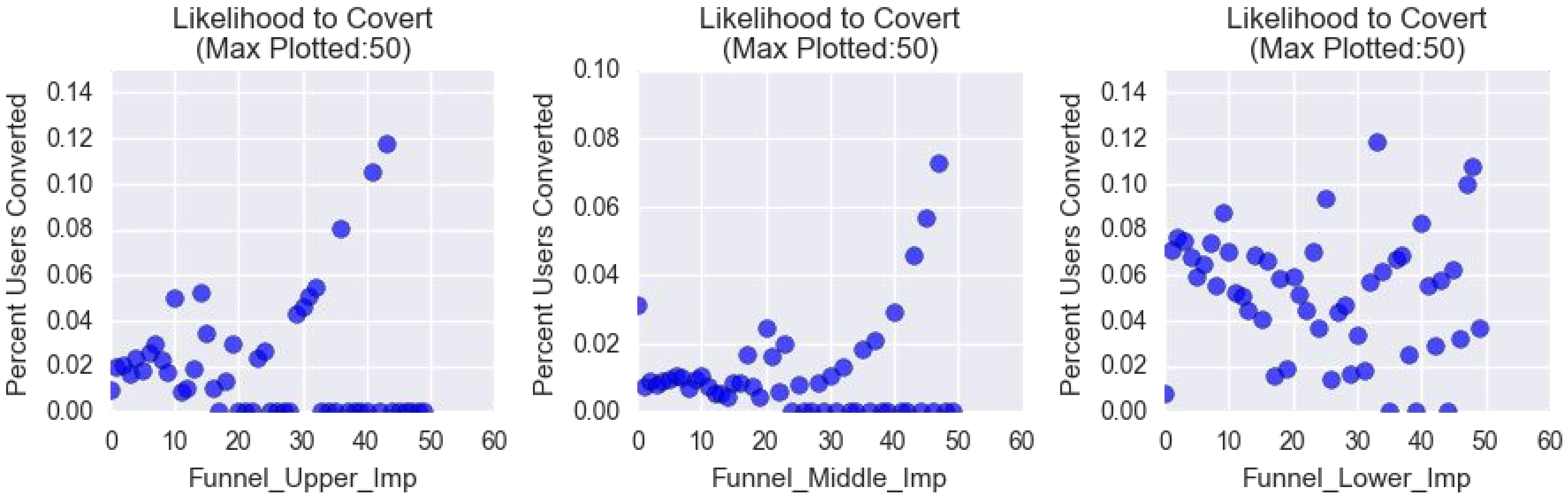


EDA - Time Difference Between First/Last Ad



- *The longer time difference has positive impact*
- *Interesting ~12hr delay has more impact than ~24hr delay*

EDA - Campaign Strategy (Funnel)



Machine Learning

Approach

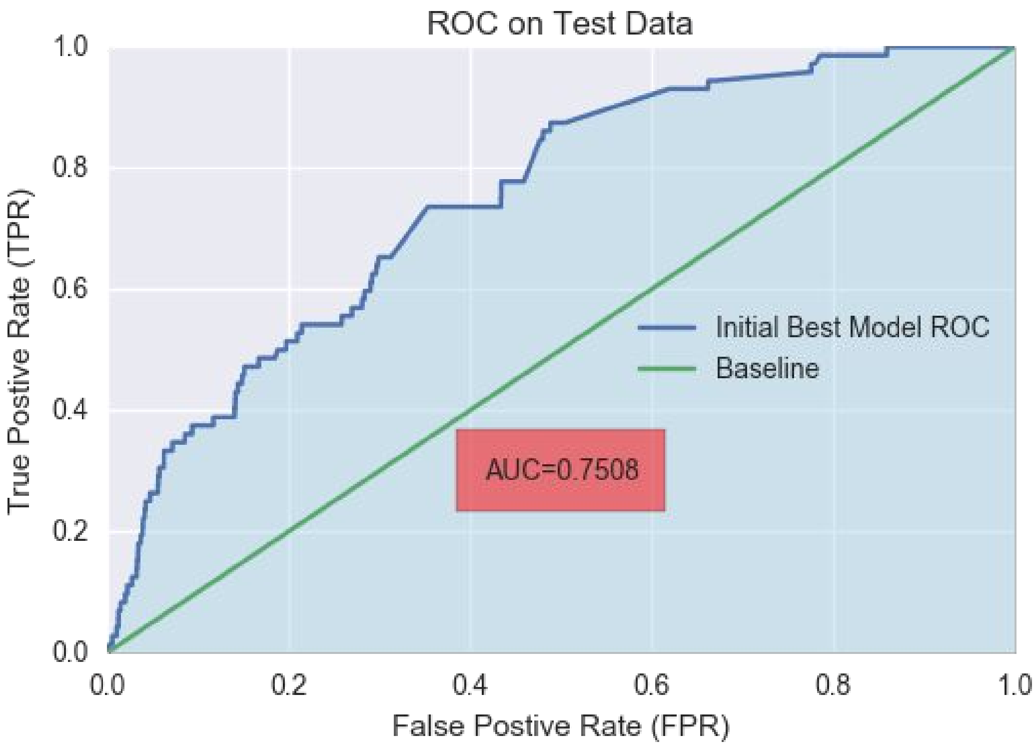
- **Model Representation**
  - **Logistic Regression** classifier
  - Use coefficients to give intuition to non-technical audience

- **Feature Selection**
  - **L1 regularization** to reduce set of highly correlated features
  - Attempt to engineering additional features

- **Model Evaluation**
  - **Area Under the Curve (AUC)**
  - Useful scoring method for situations with rare outcomes

Initial Results

L1 Logistic Regression - All Features



Features Remaining After L1 Regularization:

TimeDiff_Minutes
Funnel_Middle_Imp
Funnel_Lower_Imp
Campaign_Message_Family_Travel_Imp
Campaign_Card_Cash_Rewards_Imp
Campaign_Card_Premium_Rewards_Imp
Campaign_Card_Other_Imp
Creative_Size_320x50_Imp
Creative_Size_320x420_Imp
Device_Desktop_Imp
Device_Other_Imp
Device_Mobile_Imp
Clicks
TimeDiff_NULL_FLAG

Additional Feature Engineering

Included in Model

- **Categorial Version of Time Difference**
  - Only 1 impression (Reference Point)
  - TimeDiff 1 Day
  - TimeDiff 1 to 7 days
  - TimeDiff 7 + days

Not Included

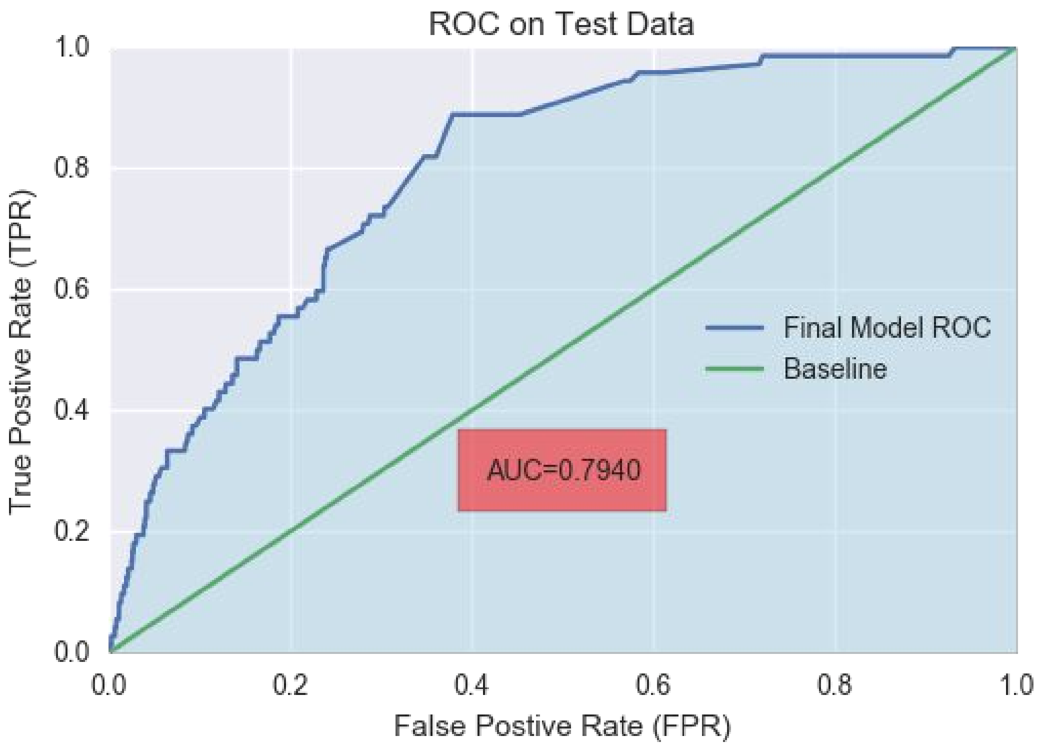
- **Funnel Halo Effect**
  - E.g. Upper_Funnel_Imp * Lower_Funnel_Imp
  - Minimal impact in AUC

- **Viewability**
  - Viewable / Measureable impressions for user
  - Regularization reduce coefficient to zero

- **Small / Medium / Large creative**
  - Group creative impressions by general size bucket
  - Minimal impact in AUC

Final Results

L2 Logistic Regression - New Representation of Time Features



Original AUC=0.7508
Final AUC=0.7940

Final Results

L2 Logistic Regression

Features	Coefficients	Odds Ratio
TimeDiff_7plus	0.074653	1.078
TimeDiff_1to7_Days	0.054029	1.056
Device_Desktop_Imp	0.041219	1.042
Funnel_Lower_Imp	0.039189	1.040
Campaign_Message_Family_Travel_Imp	0.038157	1.039
TimeDiff_One_Day	0.024293	1.025
Campaign_Card_Cash_Rewards_Imp	0.023612	1.024
Campaign_Card_Other_Imp	0.019738	1.020
Clicks	0.009595	1.010
Campaign_Card_Premium_Rewards_Imp	-0.022039	0.978
Device_Mobile_Imp	-0.026827	0.974
Device_Other_Imp	-0.043009	0.958
Funnel_Middle_Imp	-0.127673	0.880

Features with Strong Positive Influence

Features with Strong Negative Influence

Business Metrics

Cost Benefit Analysis

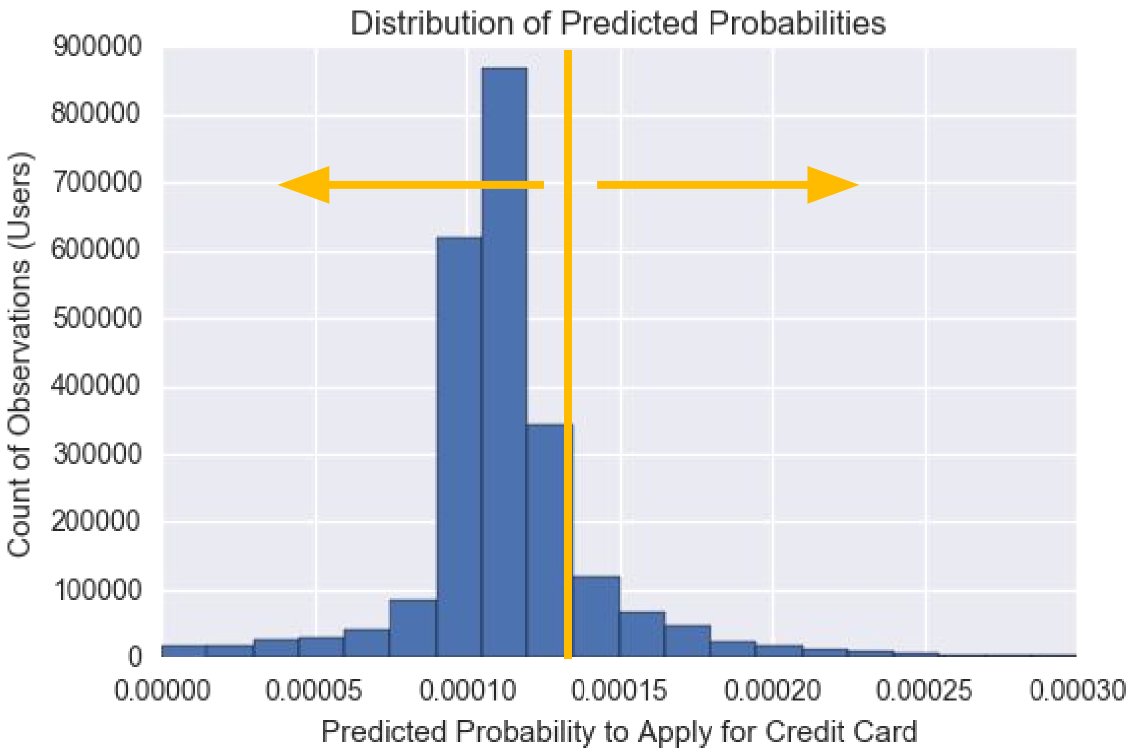
Cost-Benefit Matrix	True Class: Positive	True Class: Negative
Predicted Class Positive	Application Value - Cost of Reaching User	-Cost of Reaching User
Predicted Class Negative	0	0

Application Value = $500

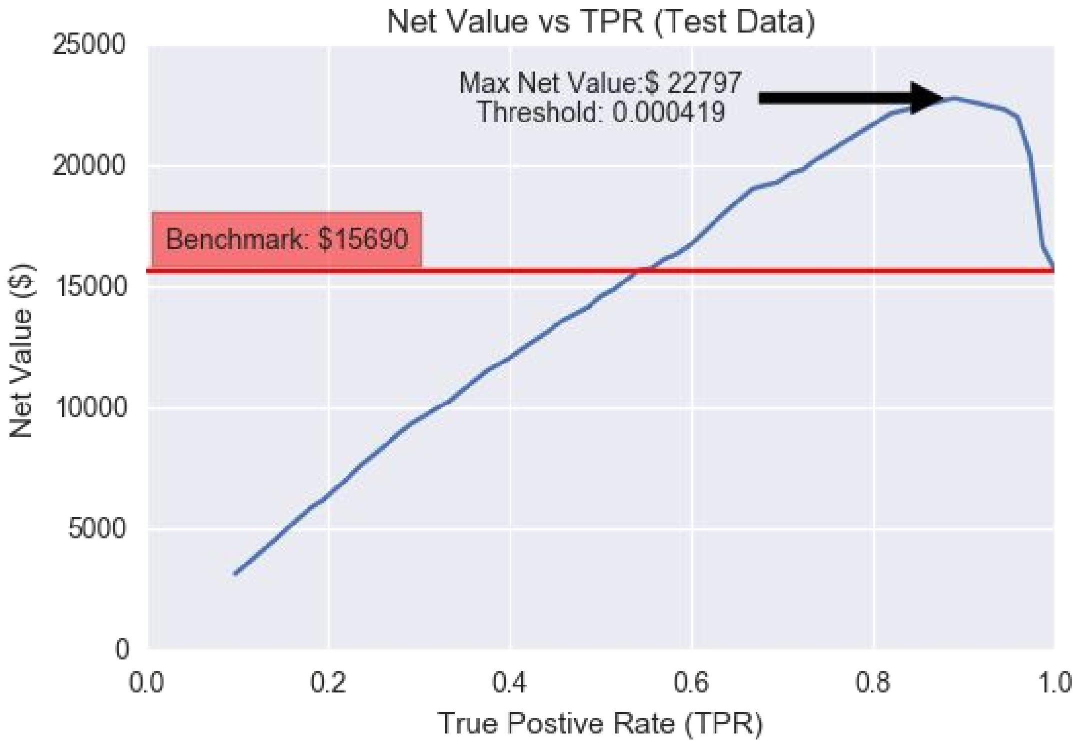
Avg Cost of Reaching User = $0.033827

Predicted Probabilities as a Signal

What is the ideal threshold of the predicted probability to signal that it is worth reaching that user?



Determine threshold to maximize **Net Value**

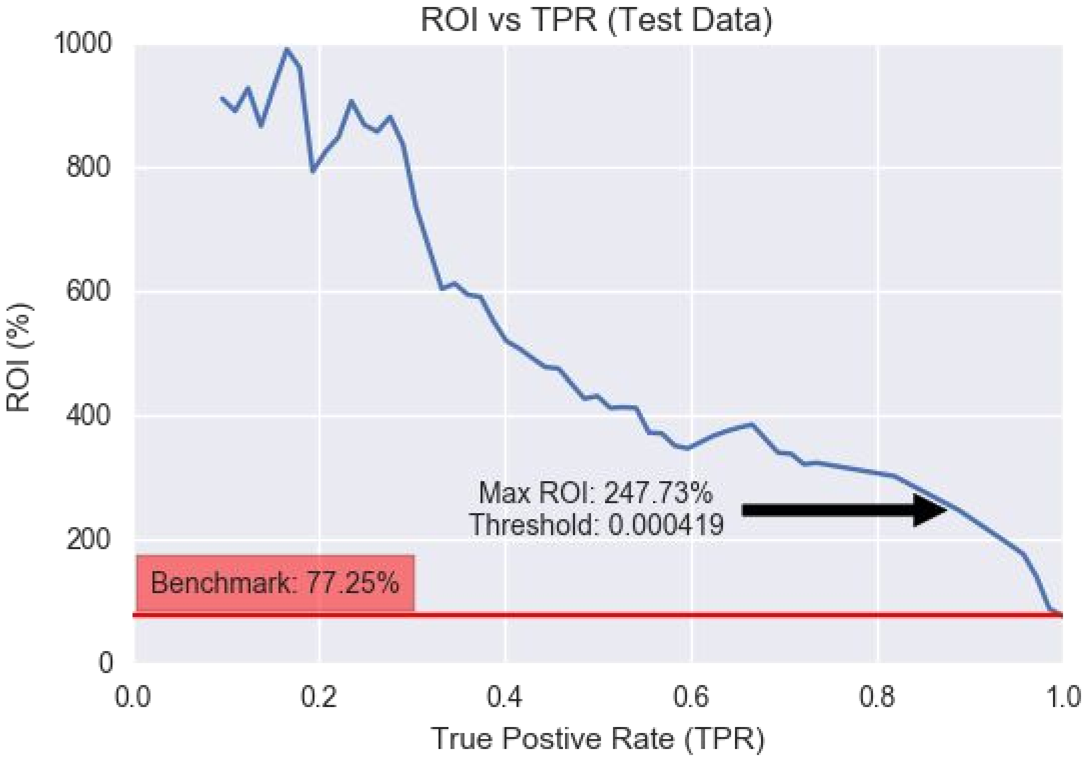


Optimal Threshold = 0.000419

Max Net Value = $22,797
Baseline = $15,690

45% improvement

Determine **ROI** at the same threshold



Optimal Threshold = 0.000419

ROI = 248%
Baseline = 77%

We could increase ROI at higher thresholds, but this would sacrifice Net Value

Conclusion and Next Steps

Conclusion

- Model provides Fair Quality Signal (AUC=0.795)
- Influential features:
  - **Positive influence**
    1. Receiving ads over several days
    2. Desktop ads
    3. Lower-funnel ads
    4. “Family Travel” ad messaging
  - **Negative influence**
    1. Middle Funnel ads
    2. “Other” Device ads
- Not in line with our hypothesis, but the outcomes are still insightful and can provide significant improvement in Net Value and ROI

Next Steps

- **Further research of results**
    - Why did the Family Travel message work so well?
    - Explore causation vs correlation:
        1. **Desktop:** users seeing mobile ads switching to desktop to apply?
        2. **Middle Funnel:** users reached with this tactic inherently less likely to apply?

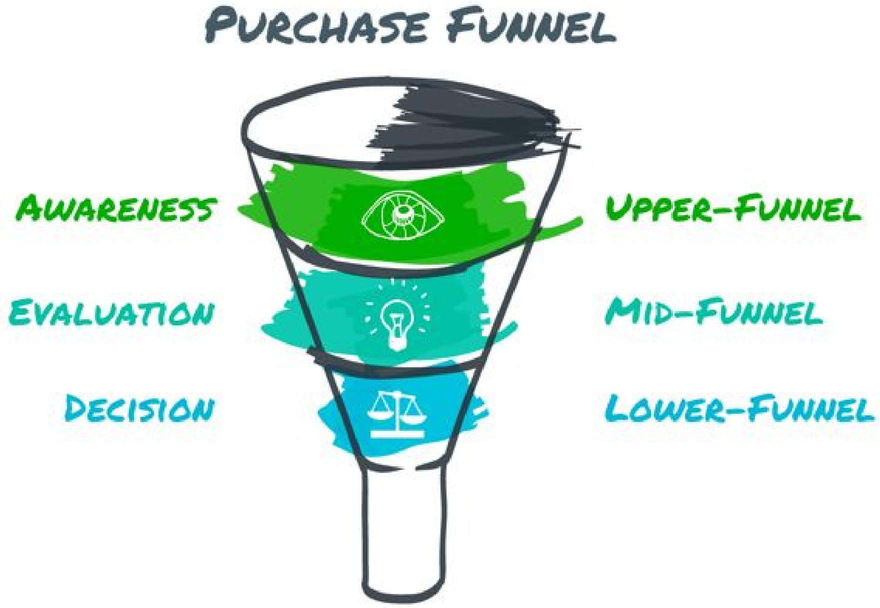
- **Additional Features**
    - User “Audience”
    - What site was the ad seen on?
        1. Site category
        2. Site quality
    - Money spent on ads for user (proxy for quality of ad placement)

Thank You!

Appendix

Need for a new approach

- High focus on “Lower-funnel” campaign strategies
- Opportunity to identify other campaign strategies to influence users to apply, e.g.:
    - Video ads
    - Interactive ads
    - Custom messaging
    - “Upper-Funnel” ads



Net Value and ROI Calculations

$$ROI = \frac{TotalValue}{TotalCost} - 1$$

$$NetValue = TotalValue - TotalCost$$

$$TotalValue = ApplicationValue * TP$$

$$TotalCost = ReachCost * (TP + FP)$$

Where TP is True Positive an FP is False Positives

Assumptions:

- **Application Value** = $500 (The average value of an online credit card application)
- **Reach Cost** = $0.033827 (The average cost of reaching a single user for our campaign)

Original Dataset Details

We have 3 separate files, one for impressions, one for clicks, and one for “activities” (e.g. credit card applications). The table below notes the dimensions we considered for this research with notes on which dimension existed in each file. For more details on DoubleClick Data Transfer files, please visit the [developer’s resources here].

Fields	Type	Impression	Click	Activity	Description
Event Time	Long	Yes	Yes	Yes	Time in microseconds since 1970-01-01 00:00:00 UTC
User ID	String	Yes	Yes	Yes	The DoubleClick cookie ID
Advertiser ID	Long	Yes	Yes	Yes	Unique ID of the advertiser
Campaign ID	Long	Yes	Yes	Yes	Unique ID of the campaign
Ad ID	Long	Yes	Yes	Yes	Unique ID of the ad placement
Rendering ID	Long	Yes	Yes	Yes	Unique ID of the creative
Placement ID	Long	Yes	Yes	Yes	Unique ID for the site page / placement where the ad ran
Browser/Platform ID	String	Yes	Yes	Yes	ID of the browser type
Active View: Eligible Impressions	Long	Yes	No	No	Whether the impression was eligible to measure viewability
Active View: Measurable Impressions	Long	Yes	No	No	Whether the impression was measurable with Active View
Active View: Viewable Impressions	Long	Yes	No	No	Whether the impression was viewable
Total Conversions	Integer	No	No	Yes	Number of Conversions
Activity ID	Long	No	No	Yes	The ID of the Floodlight tag related to the conversion event

Transformed Dataset (User-level Data) - part 1

For this research, we used SQL to transform our original dataset, which was at the impression, click, and conversion level to a single dataset at the user level. We used our original dataset to summarize the ad exposure history for each user, whether they clicked, and whether they applied for a credit card.

Field	Data Type	Type of Variable	Description
User_ID	String	N/A	Unique Identifier for each user
Impressions	Integer	Continuous	Total number of ads shown to the user
TimeDiff_Minutes	Float	Continuous	Total time in minutes between first and last impression, Null if only one impression
TimeDiff_Minutes_AVG	Float	Continuous	Average time in minutes between ad impressions, Null if only one impression
Funnel_Upper_Imp	Integer	Continuous	Total impressions from Upper Funnel campaigns
Funnel_Middle_Imp	Integer	Continuous	Total impressions from Middle Funnel campaigns
Funnel_Lower_Imp	Integer	Continuous	Total impressions from Lower Funnel campaigns
Campaign_Message_Travel_Imp	Integer	Continuous	Total impressions from ads with a "Travel" message
Campaign_Message_Service_Imp	Integer	Continuous	Total impressions from ads with a "Service" message
Campaign_Message_Family_Travel_Imp	Integer	Continuous	Total impressions from ads with a "Family Travel" message
Campaign_Card_Cash_Rewards_Imp	Integer	Continuous	Total impressions from ads with a "Cash Rewards" message
Campaign_Card_Premium_Rewards_Imp	Integer	Continuous	Total impressions from ads with a "Premium Rewards" message
Campaign_Card_Other_Imp	Integer	Continuous	Total impressions from ads with a "Other Card" message
Creative_Type_Display_Imp	Integer	Continuous	Total impressions from creative type "Display"
Creative_Type_TrueView_Imp	Integer	Continuous	Total impressions from creative type "TrueView"
Creative_Type_RichMediaExpanding_Imp	Integer	Continuous	Total impressions from creative type "Rich Media Expanding"

Transformed Dataset (User-level Data) - part 2

For this research, we used SQL to transform our original dataset, which was at the impression, click, and conversion level to a single dataset at the user level. We used our original dataset to summarize the ad exposure history for each user, whether they clicked, and whether they applied for a credit card.

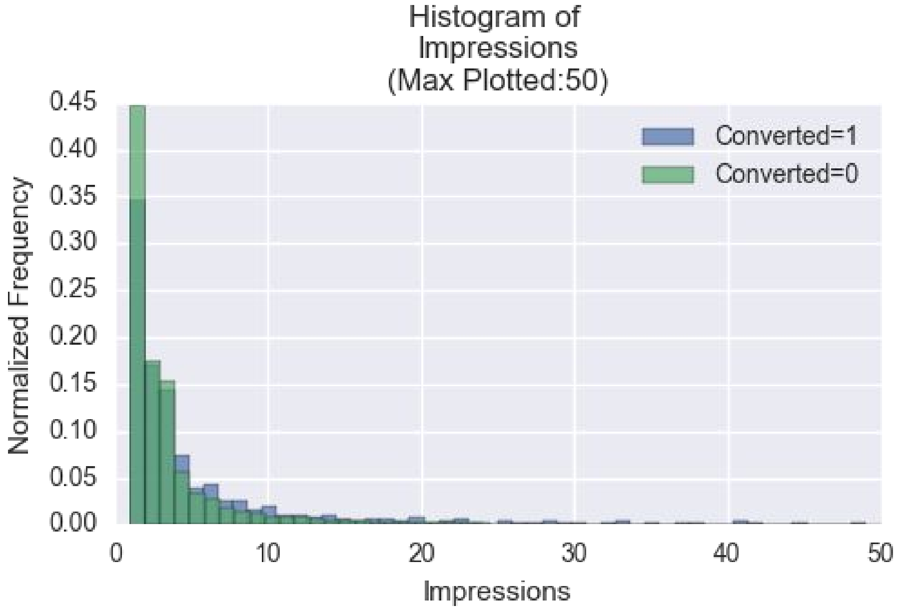
Field	Data Type	Type of Variable	Description
Creative_Type_RichMedia_Imp	Integer	Continuous	Total impressions from creative type "Rich Media Display"
Creative_Size_728x90_Imp	Integer	Continuous	Total impressions from creative size 728x90
Creative_Size_300x600_Imp	Integer	Continuous	Total impressions from creative size 300x600
Creative_Size_300x250_Imp	Integer	Continuous	Total impressions from creative size 300x250
Creative_Size_160x600_Imp	Integer	Continuous	Total impressions from creative size 160x600
Creative_Size_468x60_Imp	Integer	Continuous	Total impressions from creative size 468x60
Creative_Size_300x50_Imp	Integer	Continuous	Total impressions from creative size 300x50
Creative_Size_320x50_Imp	Integer	Continuous	Total impressions from creative size 320x50
Creative_Size_320x420_Imp	Integer	Continuous	Total impressions from creative size 320x420
Creative_Size_480x320_Imp	Integer	Continuous	Total impressions from creative size 480x320
Creative_Size_320x480_Imp	Integer	Continuous	Total impressions from creative size 320x480
Creative_Size_Uknown_Imp	Integer	Continuous	Total impressions from unknown creative size
Device_Desktop_Imp	Integer	Continuous	Total impressions from desktops
Device_Other_Imp	Integer	Continuous	Total impressions from mobile devices
Device_Mobile_Imp	Integer	Continuous	Total impressions from other devices (e.g. gaming consoles)

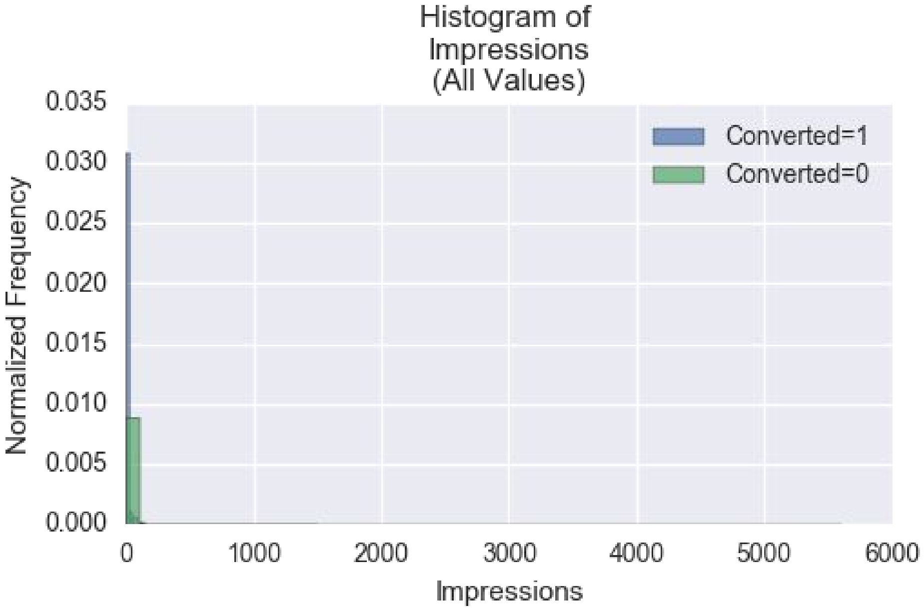
Transformed Dataset (User-level Data) - part 3

For this research, we used SQL to transform our original dataset, which was at the impression, click, and conversion level to a single dataset at the user level. We used our original dataset to summarize the ad exposure history for each user, whether they clicked, and whether they applied for a credit card.

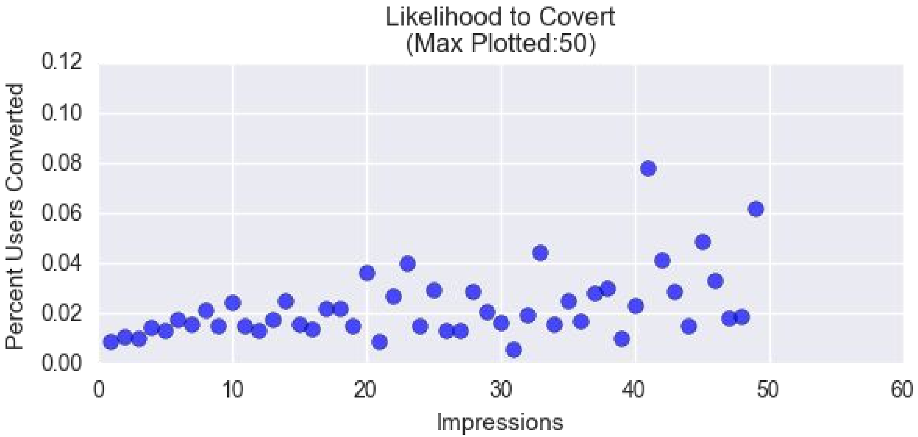
Field	Data Type	Type of Variable	Description
Active_View_Eligible_Impressions	Integer	Continuous	Total impressions that were eligible for viewability measurement
Active_View_Measurable_Impressions	Integer	Continuous	Total impressions that were measureable
Active_View_Viewable_Impressions	Integer	Continuous	Total impressions that were viewable
Clicks	Float	Continuous	Total ad clicks
Conversions	Float	Continuous	Total conversions, i.e. online credit card applications

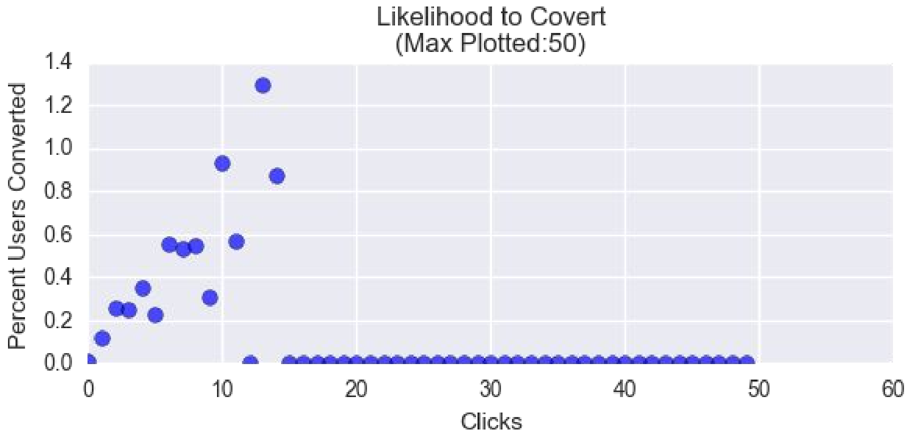
EDA - Impression Distribution



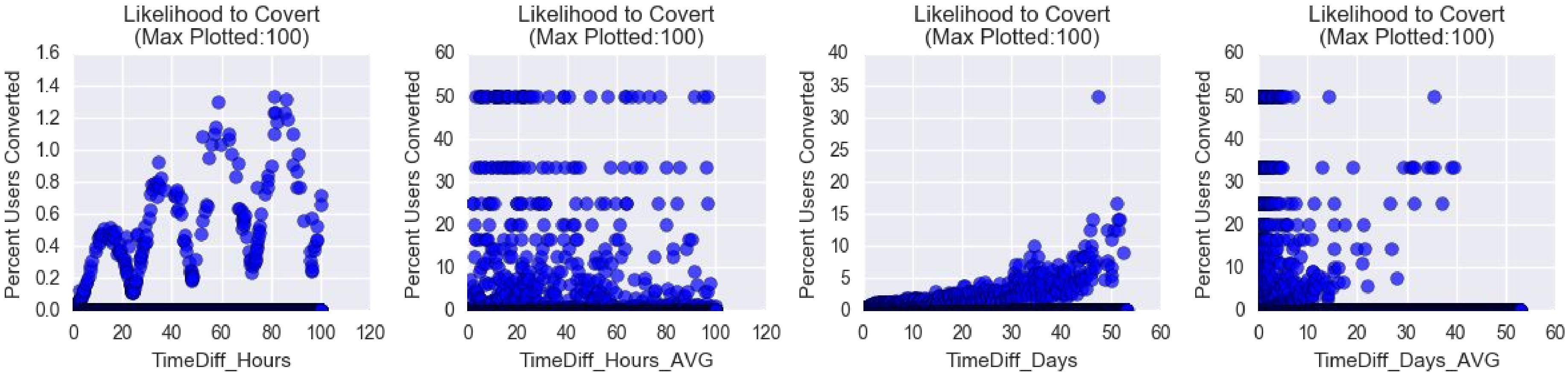


EDA - Impressions and Clicks

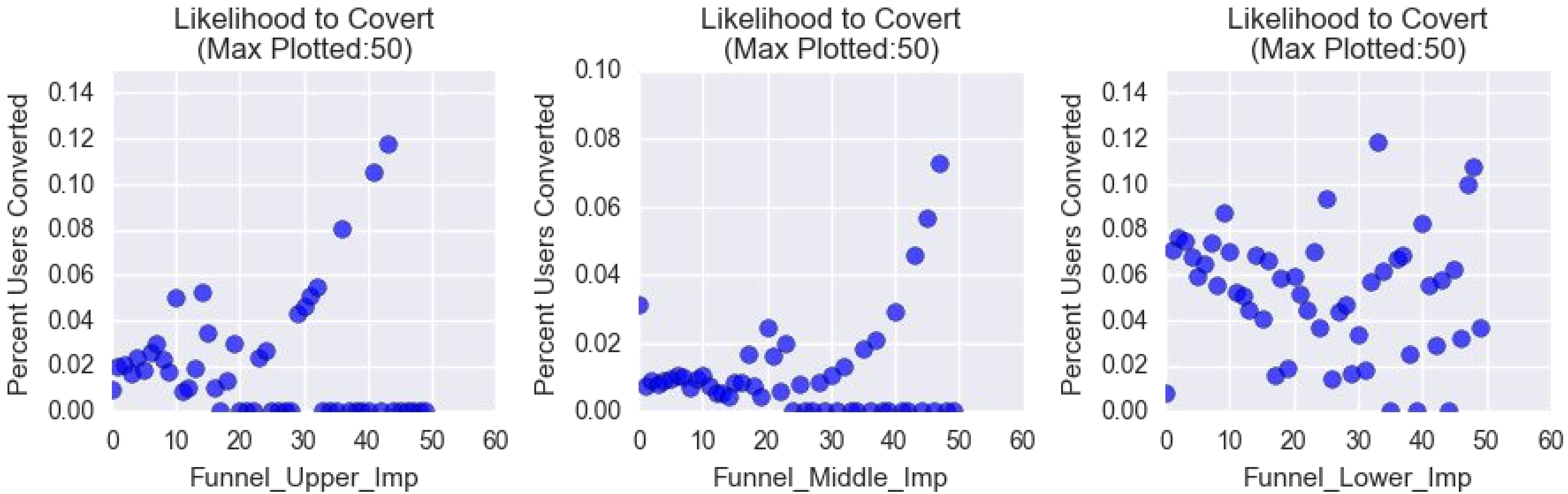




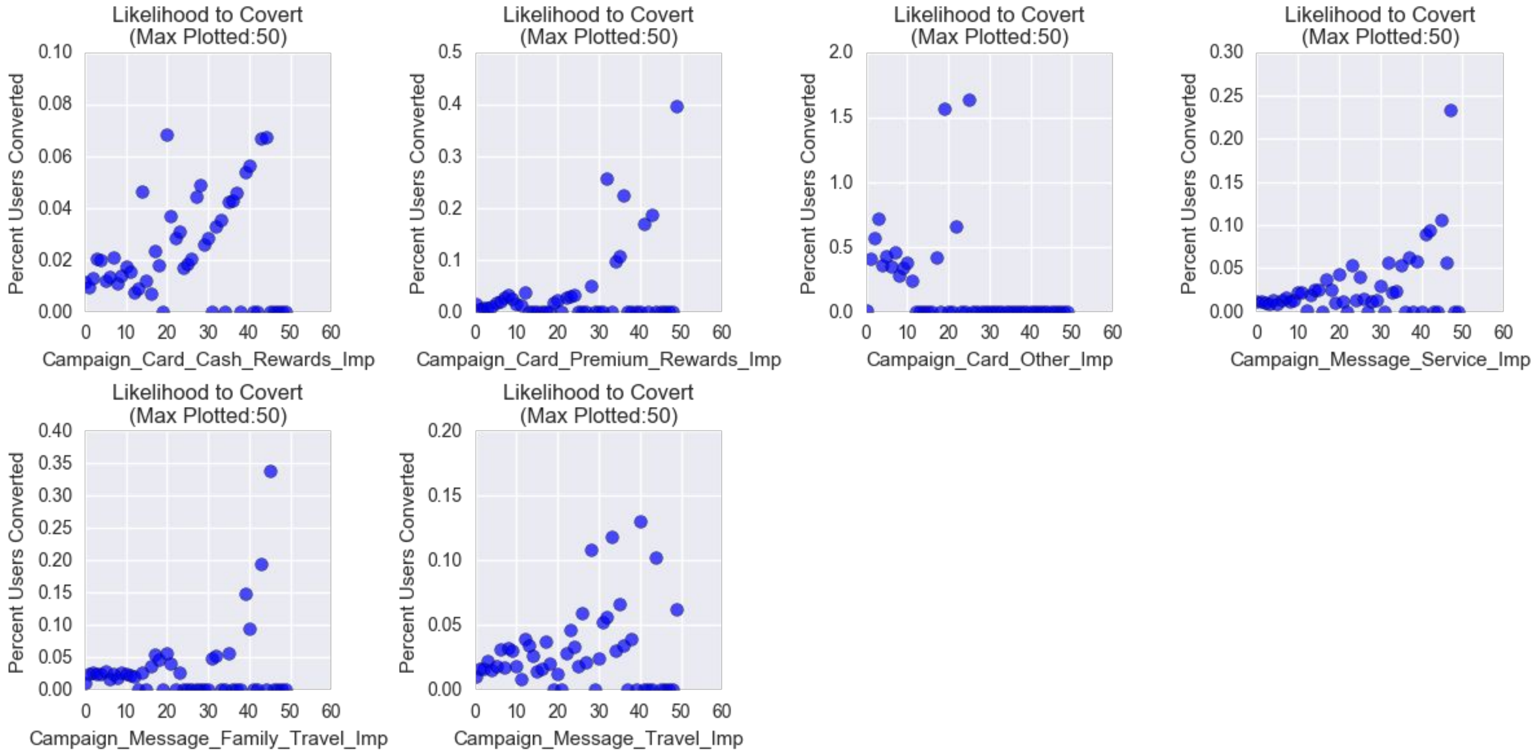
EDA - Time Difference Between Ads



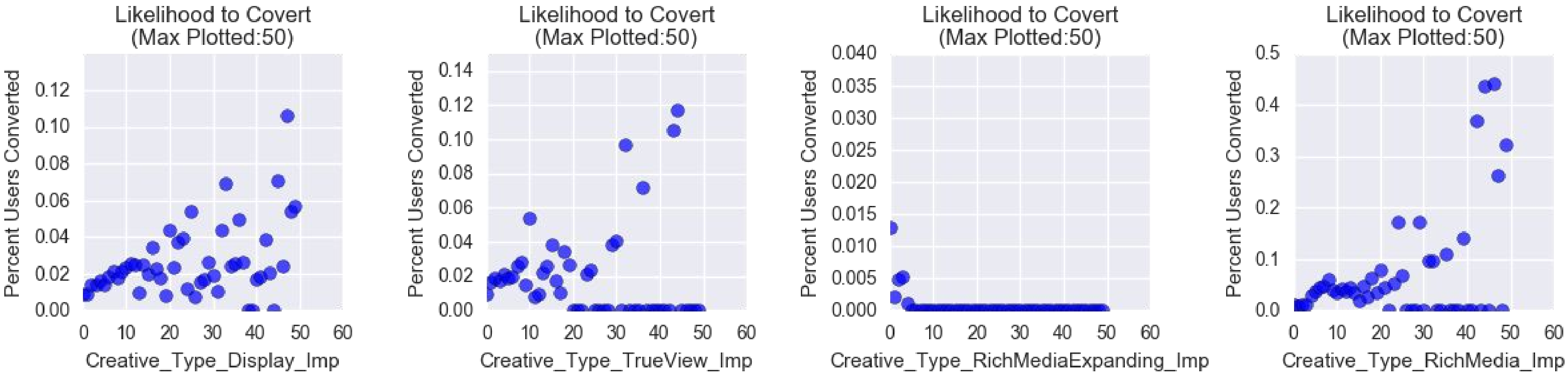
EDA - Campaign Strategy (Part of Funnel)



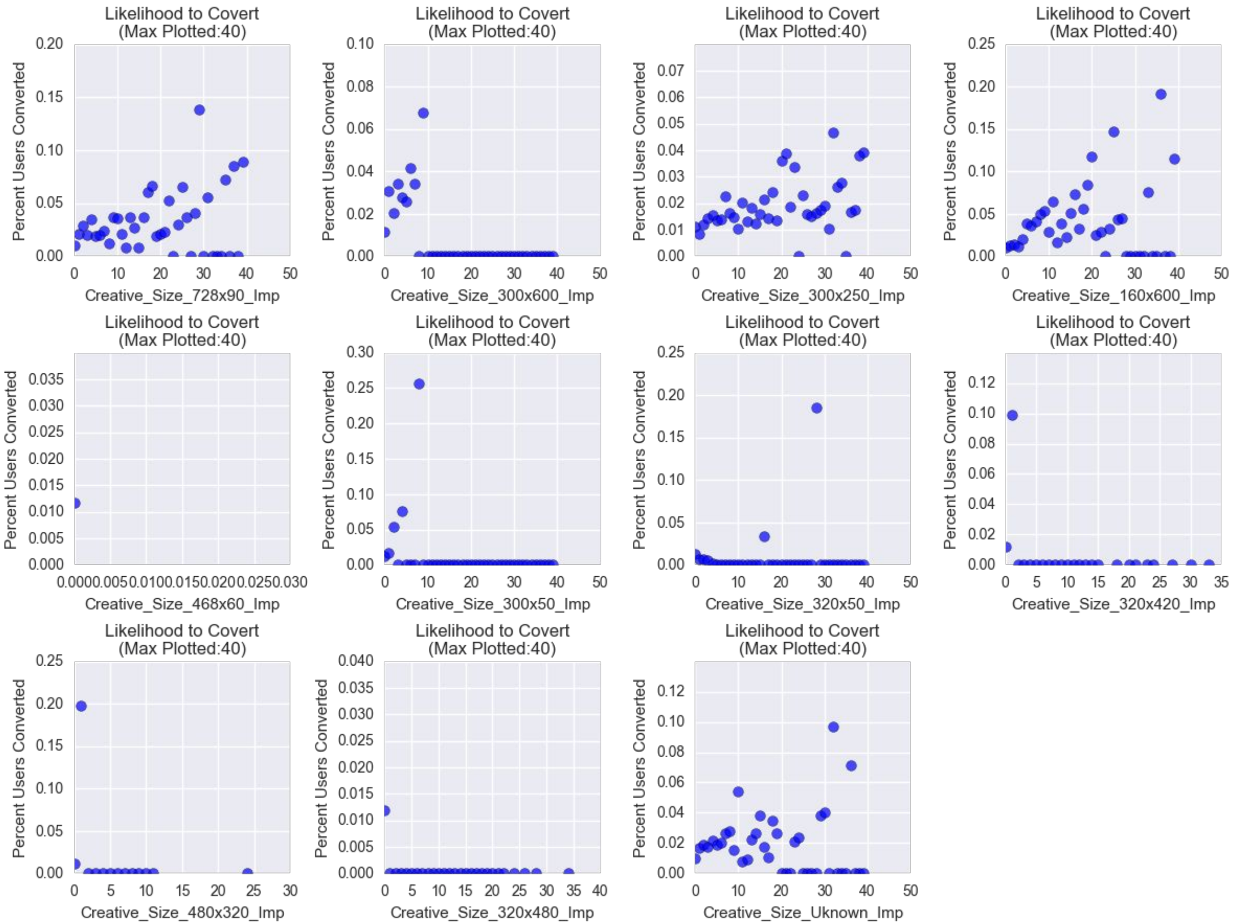
EDA - Creative Message/Card Type



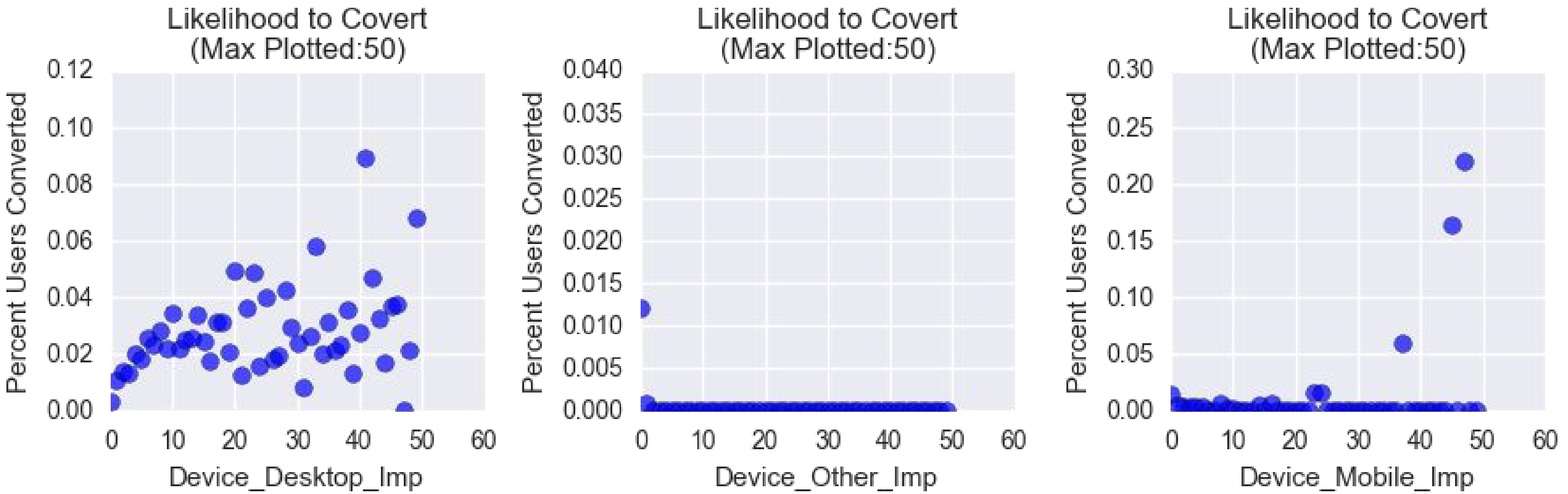
EDA - Creative Type



EDA - Creative Size



EDA - Device Type



EDA - Viewability