Clio

Data Analysis

Geoffrey Shmigelsky
Tuesday, Oct 29, 2019
Time Spent: Six Hours

Executive Summary

Clio is interested in understanding and increasing conversion rates, specifically:

"How do we increase paid conversion rates in the first 14 days?"

The purpose of this analysis is to identify where to optimize best the prospect of conversion based on eight trackable features; four are web-based tracking; four are user milestones. Currently, 1 in 9 prospects convert in 14 days.

In the supplied dataset, "Time to Conversion" and "Conversion Value" input features define a Conversion numerically. For this analysis, a conversion is either a "Yes" or "No" based on a value existing in either field.

Recommendation

The hypothesis: If a user is prepared to perform a financial transaction on the Clio platform, they are much more likely to convert a client.

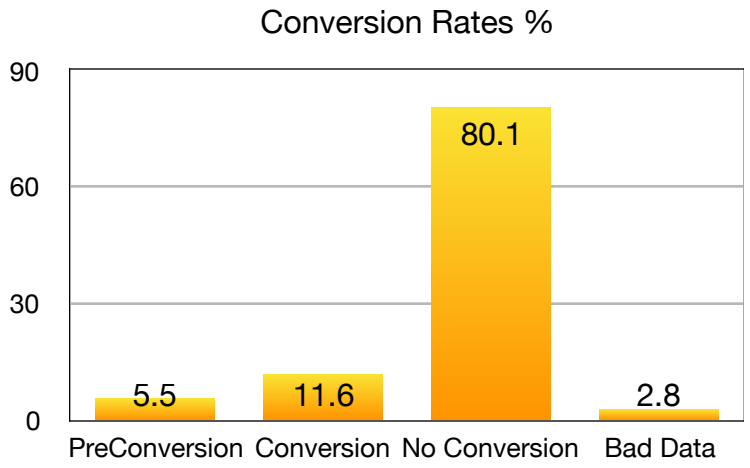
If the Clio platform website signs up 1,000 clients, the website should offer 50% of its users the control interface and 50% the experimental interface - for one month. The difference in conversions between the two would quantifiably measure the effectiveness of the change.

I suggest optimizing the process for milestones "Time First Entry" to the "First Bill." Currently, based on the "First Bill" for a successful conversion, 62% complete this step from the previous step. For failed conversions, only 48% finished the step.

The absolute numbers tell a different story; 114 prospects are lost here. For comparison, there are only a total of 113 valid conversions, all else being equal.

Describe the Data

The training dataset consists of 1,000 rows, with 171 successful conversions and 829 unsuccessful.



Several conversions occurred before the first matter, 58 in total, or 1 in 3. That is, 1 in 3 customers became conversions before any milestones and therefore removed from the training set. There are also 28 failed conversions removed from the training set as they were logically not possible.

Statistics for the remaining training data:

	count	count %	mean	std	min	max
time_to_first_matter	379	41%	81816	198154	77	1130506
time_to_first_time_entry	313	34%	87473	195931	15	1207301
time_to_first_bill	164	18%	164917	268088	217	1204219
time_to_second_user	84	9%	226122	330304	101	1198713
page_views_in_first_hour	914	100%	20	23	1	149
page_views_in_first_day	914	100%	37	68	1	673
page_views_in_first_7_days	914	100%	66	147	1	1813
page_views_in_first_14_days	914	100%	86	197	1	2117

From this data, it can be seen that 1 in 7 prospects are becoming clients of Clio in the first 14 days, which means 6 in 7 are not. Even a modest conversion improvement on failed opportunities would make a material difference.

Feature Ranking

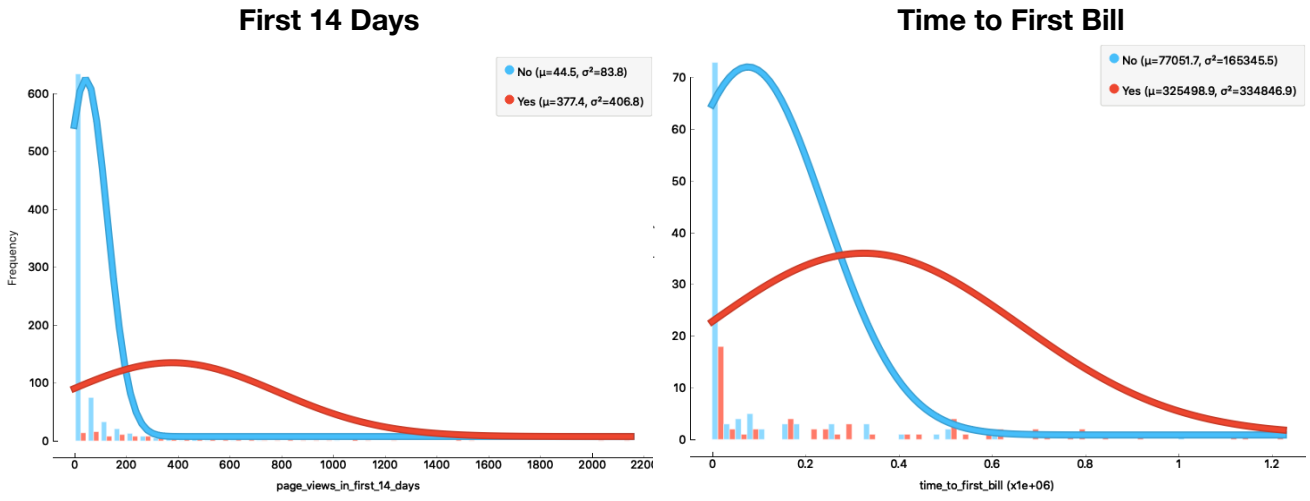
Ranking the features illustrates that page views in the first 14 days are the most telling factor. Which makes sense as the more a client uses the service, the more likely they are to convert.

What is more interesting the time to the first bill; even though it mostly an empty feature set, it is a stronger indicator of conversion relative to other milestones.

Feature	Gain ratio	Gini
first_14_days	**0.126**	**0.102**
first_7_days	0.096	0.080
first_day	0.049	0.040
first_hour	0.022	0.018
time_to_first_bill	**0.019**	**0.024**
time_to_first_time_entry	0.018	0.023
time_to_first_matter	0.015	0.019
time_to_second_user	0.003	0.004

Distributions

Below are two distributions for the conversion classifications, first_14_days and time_to_first_bill. Red indicates a **Yes** conversion; No is **Blue**.



Note how the distribution shifts between Yes and No classes; the area under the curves differ. It indicates a difference in the correlations between these two features and the classification. The other distributions for the remaining six input features are not as insightful as their differences are smaller.

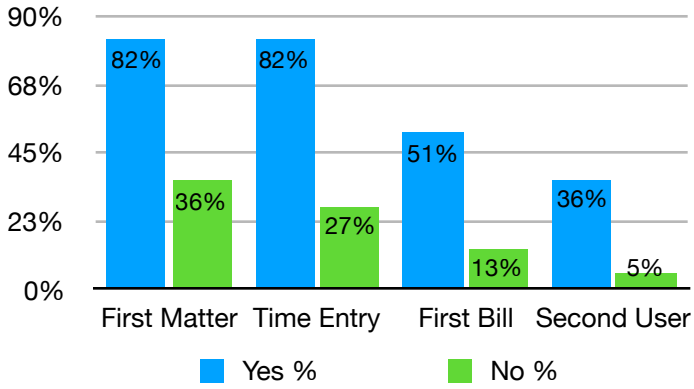
MileStones Features

If you take it one step further and ask, where are conversions being lost in the milestones? I am assuming that the milestones following a linear progression, initial analysis indicates this is mostly correct.

Milestone Transitions

Absolute	Yes	Yes %	No	No %
First Matter	**93**	82%	286	36%
Time Entry	**93**	82%	220	27%
First Bill	58	51%	106	13%
Second User	41	36%	43	5%
Totals	113		801	

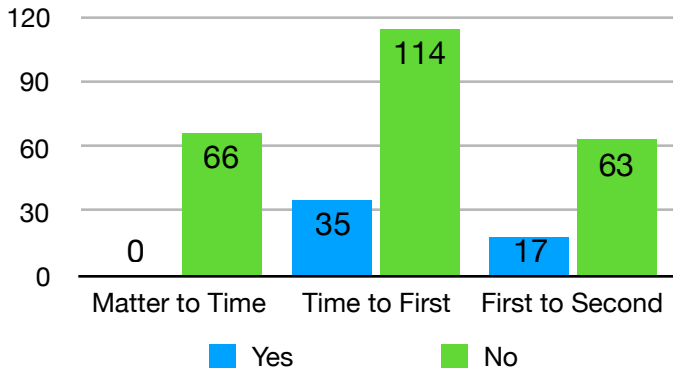
Milestone Transitions %



MileStone Change

	Yes	Yes %	No	No %
Matter to Time	0	1%	66	23%
Time to First	35	38%	**114**	52%
First to Second	17	29%	63	59%
Totals	113		801	

MileStone Transitions Change - Losses



These charts and graphs illustrate the percentages of Yes and No conversions per milestone. For example, for Yes conversions, 100% of lawyers who create a First Matter will create a Time Entry - 93 in both cases. That is not the case for No conversions; the number drops from 286 to 220 of total No prospects - a change of 23%.

The key takeaway is that the most significant drop in No prospects occurs at the transition fro Time Entry to First bill; 114 lawyers fail to bill for their time entry - a huge loss. Half the possible prospects fail to complete this step.

Incidentally, only 1 in 3 prospects who do not convert ever enter their first matter and that is a big deal in itself.

Time Features

With feature engineering, additional information can be extracted from the four time inputs. First, the times are normalized the be hourly representations. Second the different between the four points define three additional values representing the net change.

Although it is not part of this project, I would suggest defining a slope or quadratic equation for the three points, as they could be further inputs into the model.

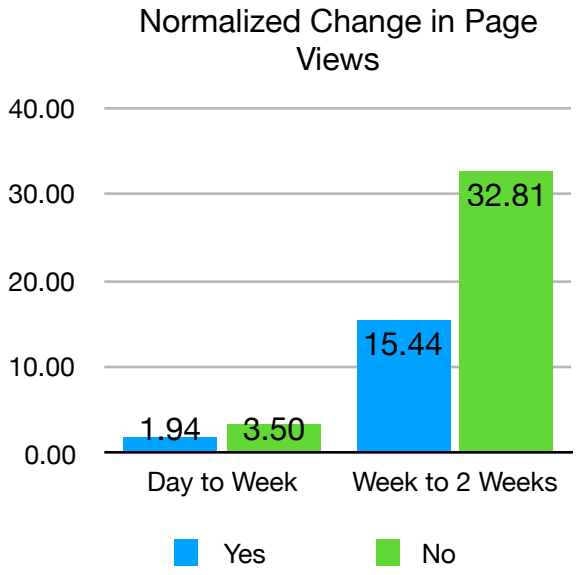
Hourly Page Views

Page Views	Yes	No
First Hour	36.150	17.649
First Day	13.300	4.108
7 Days	1.520	0.236
14 Days	0.760	0.118
Totals	113	801

Hourly Page View Changes

Page Views	Yes	Ratio	No	Ratio
Hour to Day	22.840		13.540	
Day to Week	11.780	1.94	3.872	3.50
Week to 2 Weeks	0.763	15.44	0.118	32.81
Totals			801	

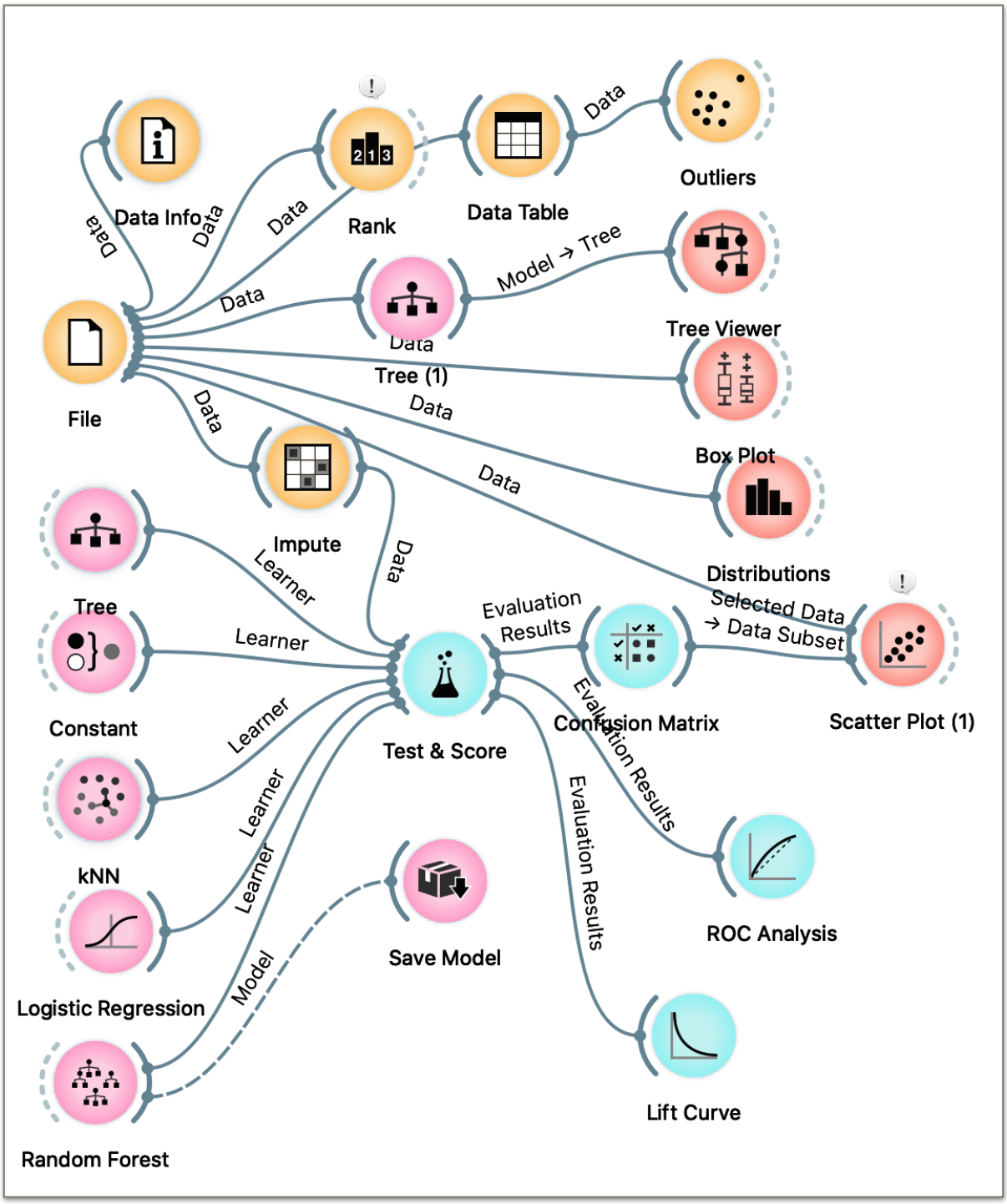
The table on the left illustrates the drop in hourly page views over time frames. Note how the drop is more substantial for prospects who fail to convert.



Again you can see that the drop off for No vs Yes is much more substantial. When a project fails to convert, it can be seen early on, as their engagement is two-thirds less on the first day.

In my opinion, page views are a consequence, not a catalyst for prospect conversion. It can tell what is happening and when, but not why.

Modelling



The tool used for modelling is the Scikit-learn GUI wrapper Orange. It is useful to explain the process and graphs. The website:

https://orange.biolab.si/download

Model Comparison

The purpose of the model is to predict the outcome of an early application engagement accurately. Recall the results is a “Yes” or “No.”

Test and Score

Model	AUC	CA	F1	Precision	Recall
Logistic Regression	0.88	0.91	0.91	0.91	0.91
Random Forest	0.90	0.91	0.90	0.90	0.91
kNN	0.80	0.88	0.87	0.86	0.88
Tree	0.60	0.87	0.86	0.86	0.87
Constant	0.49	0.88	0.82	0.77	0.88

The above chart compares different machine learning models and their performance. The most important numbers are Area Under the Curve (AUC) Classification (CA) and F1 Score. The bottom line is that the models can predict, with a 91% classification accuracy, what will happen to a prospect. The best model overall is Random Forest.

However, there is one issue to be aware of; there is an imbalance in the Yes and No classification count. Usually, you would try to balance the classes counts. However, as the goal is to optimize the No prospect process, the imbalance is left as-is for illustrative purposes.

Confusion Matrix - Proportion of Actual

	No	Yes	Sum
No	96.8%	3.2%	801
Yes	52.2%	47.8%	113
Sum	834	80	91

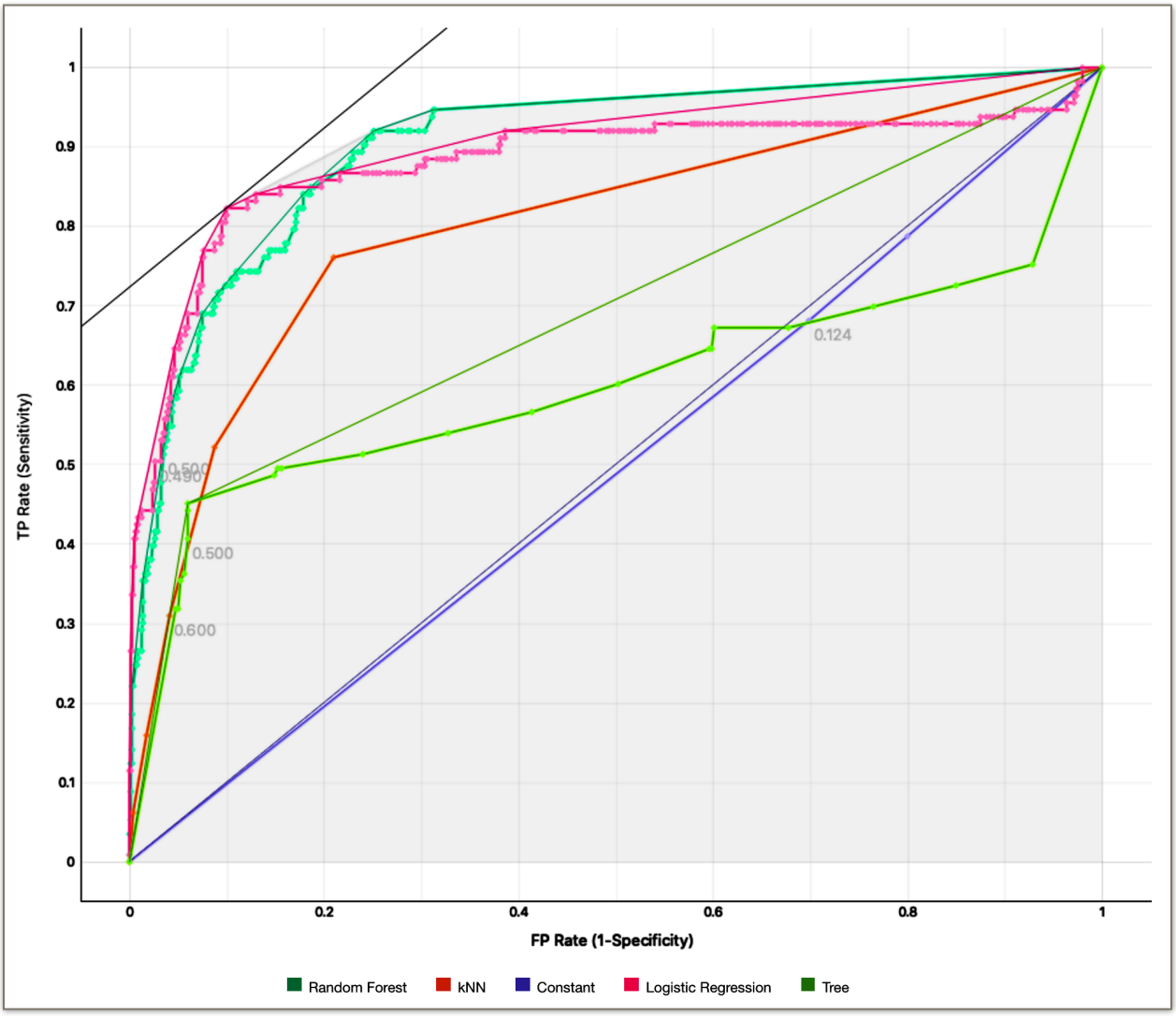
Confusion Matrix - Proportion of Predicted

	No	Yes	Sum
No	92.9%	32.5%	801
Yes	7.1%	67.5%	113
Sum	834	80	91

Confusion matrixes show how a model can predict each class. For Random Forest, its accuracy 93 percent or better for the No case.

Model Comparison

From a modelling point of view, the Receiver Operating Characteristic Curve (ROC) is a better representation of model performance.



The chart demonstrates the slight difference in sensitivity and specificity between Logistic Regression and Random Forest.

In summary, a Random Forest will be sufficient for modelling purposes. However, the classes should be balanced to prevent the model from favouring one class over another.