Predictive Scoring Tutorial

This article supports Audience Studio - Legacy.

The goal of this tutorial is to perform a churn prediction on public data from a US telecom. This tutorial uses a data set that consists of the churn history of phone numbers from the book “Discovering Knowledge in Data: An Introduction to Data Mining.”

- Create Batch Segments Which Represent the Churn Behavior
- Configure Predictive Scoring
- Learning from the Population, and Assigning Predictive Scores to Customers
- Review the Data on the Dashboard
- Create a New Batch Segment Based on the Predictive Scores
- Understand Predictive Model and Tune It to Achieve Higher Accuracy

This is a tutorial overview and does not include steps or screen captures for a complete scenario. Assume the data is already imported to Treasure Data as a table:

01_demo_db / prediction_sample

QUERY

Records Columns Details

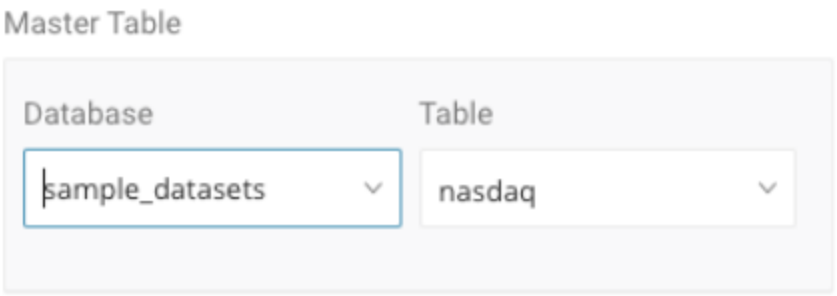
40 of 1,000,000 records • Jump to time

	# time	Ab phone	.# eve_mins	# intl_calls	Ab int_l_plan	.# night_charge
1	1473055233	399-5763	165.7	6	no	11.97
2	1473055233	340-9449	225.7	4	yes	11.49
3	1473055233	363-1123	92.0	17	no	10.12
4	1473055233	361-2170	283.3	5	no	8.47
5	1473055233	406-6304	144.4	4	no	11.81
6	1473055233	392-5090	143.7	5	no	8.61
7	1473055233	373-7713	233.7	4	no	5.94
8	1473055233	392-1376	114.5	4	no	8.02
9	1473055233	390-6388	219.1	9	no	9.91
10	1473055233	379-7290	188.5	8	yes	9.5
11	1473055233	397-9928	300.5	4	no	8.66
12	1473055233	378-6924	209.9	6	no	12.64
13	1473055233	411-5677	196.4	4	yes	5.4
14	1473055233	400-2738	85.0	8	no	9.45
15	1473055233	409-1856	265.5	2	no	8.12
16	1473055233	362-5899	249.3	3	no	10.22
17	1473055233	377-1164	197.7	4	no	8.72
18	1473055233	368-8555	116.9	4	no	10.95
19	1473055233	347-6812	284.8	7	no	8.05
20	1473055233	348-3830	189.7	5	no	9.96
21	1473055233	414-4276	215.5	6	no	12.56

The table has 1,000,000 records (such as customers, phone numbers; 1 record = 1 phone number), and each profile has 20 attributes (such as day calls, account length and international plan) and 1 label column “Churn” (`True.` or `False.`).

The goal is to create a predictive model that determines the customers who are likely to churn in the near future.

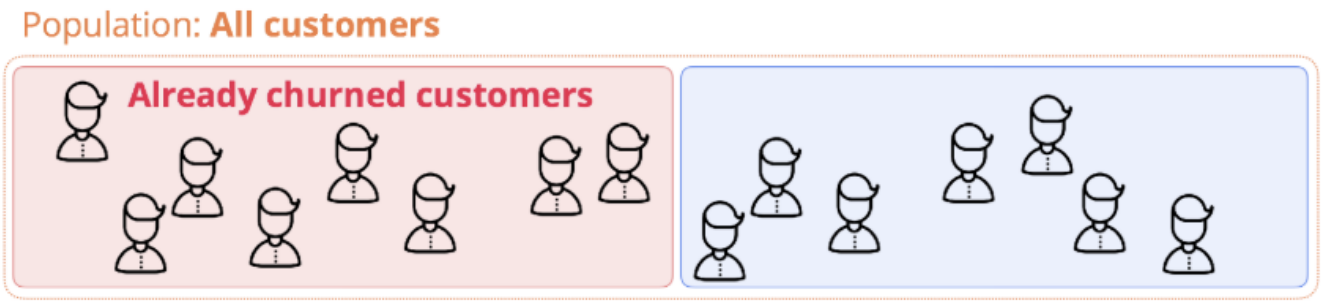
1. Create a master segment based on the data. Because the data is quite simple, you can create a master segment by directly using the table as a master table:
2. Click Run to generate the master segment data.



You might need to preprocess your data to create a reasonable master segment with informative attributes.

Create Batch Segments Which Represent the Churn Behavior

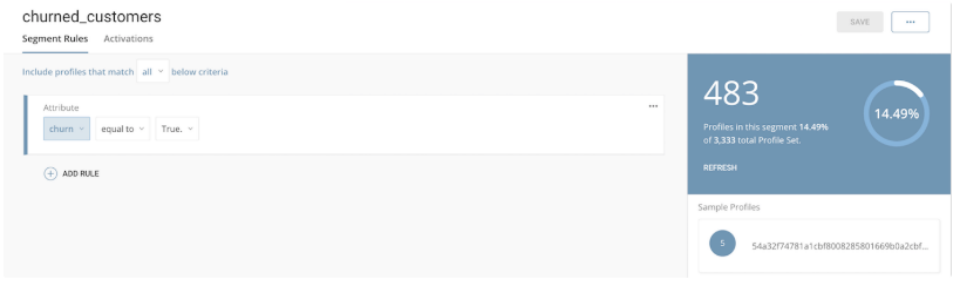
Define batch segments representing a churn prediction. In this example, the goal is to put predictive scores to customers who have not churned yet. See Predicting Customer Behavior. The separation between **population**, **positive samples,** and **scoring target** can be illustrated as follows:



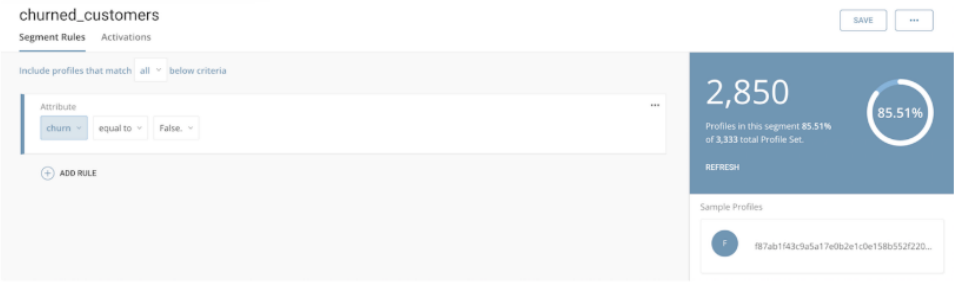
Create segments. Separate segments mean that:

1. A predictive model is built based on the full master segment, and the model represents characteristics of customers who are in the **positive samples** segment.
2. In a scoring step, only active customers get a predictive score according to their possibility of future churn.

- **Positive samples**



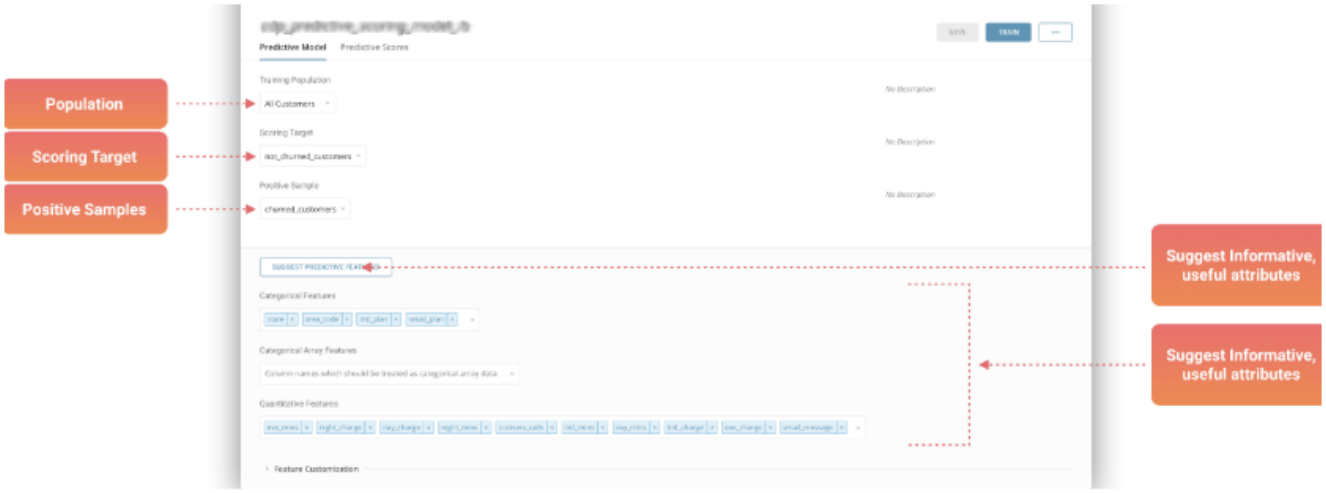
- **Scoring target**



Configure Predictive Scoring

After the segments are defined, you are ready to implement predictive scoring.

Specify the dependent segments and attributes used for prediction in Predictive Scoring:



Choosing a subset of attributes is a part of feature engineering in the context of machine learning. Data scientists generally spend significant amounts of time to find an appropriate feature set.

To allow non-experts to choose reasonable attributes, you can use Treasure Data's feature guess function. Click Suggest Predictive Features to see a suggested set of attributes that you can use to make a reasonable prediction on your master segment and segments. See How Feature Guess Works.

Selected columns are categorized into the following types:

Categorical Features	• Attributes which are not meaningful as a numeric value such as gender, day of week, group etc.
Categorical Array Features	• Array column on TD which can be treated as single categorical information such as td_affinity_categories generated by the content affinity engine, list of games played before etc.
Quantitative Features	• Numeric values such as age, price, frequency etc.

You can add and remove columns.

Learning from the Population, and Assigning Predictive Scores to Customers

Ultimately, what you need to do for predictive scoring is:

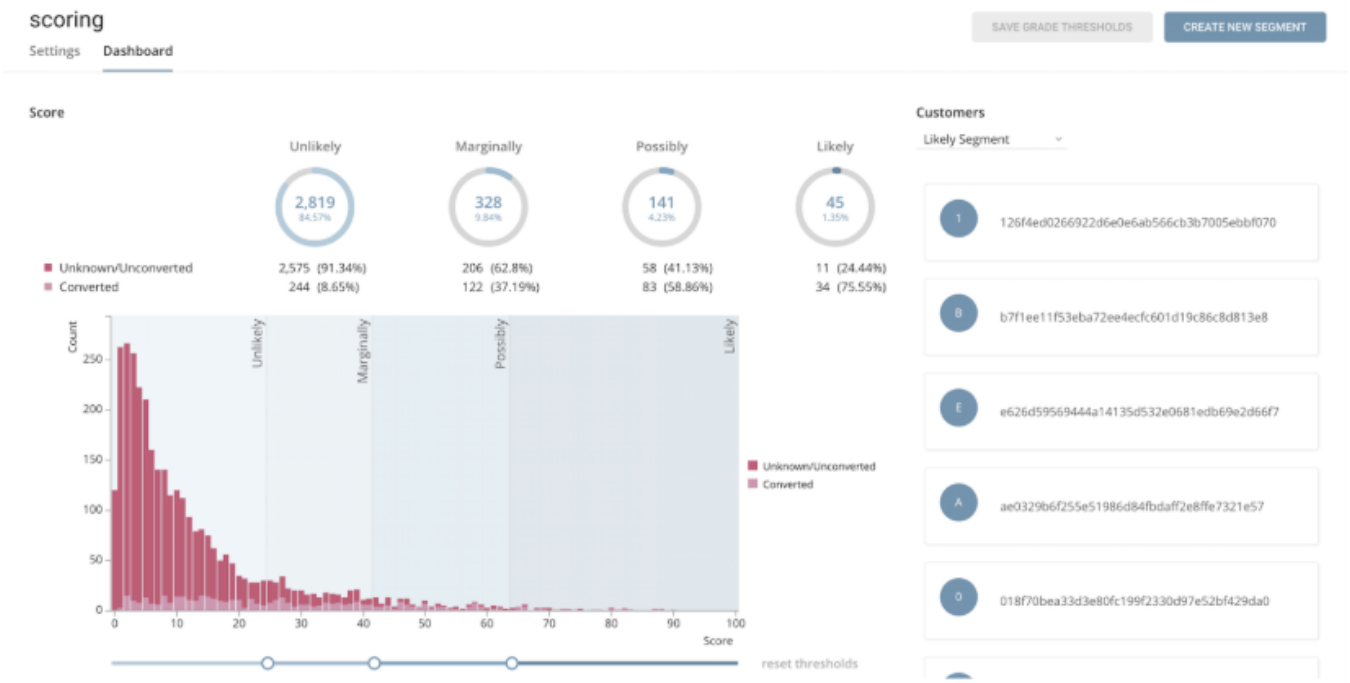
1. Click Run on the Predictive Scoring view (~10min)
2. Click Run on the Master Segment view (~7min)

Each run operation internally executes Treasure Workflow. You cannot edit these internal workflows.

- The first workflow learns characteristics of customers (profiles) who are in the master segment, and builds a predictive model,.
- The second workflow re-generates the master segment and assigns a predictive score to profiles in the scoring target segment at the time when you click Run.

Review the Data on the Dashboard

After the master segment is successfully re-generated, review your dashboard:



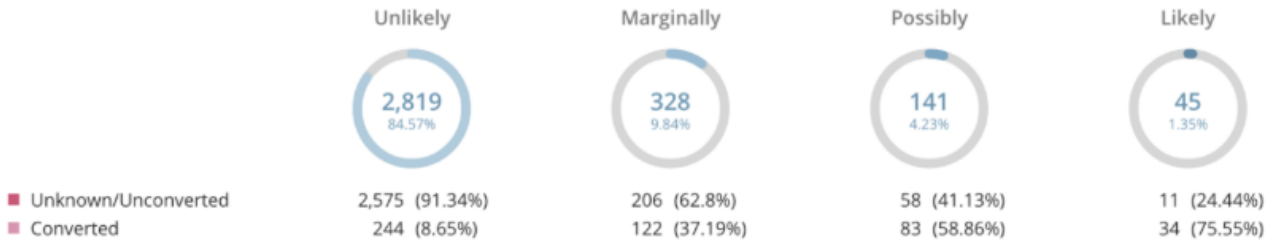
The histogram shows the distribution of predictive scores. The horizontal axis corresponds to predictive score distributed from 0 to 100. The vertical axis indicates the number of scored profiles (customers). The different colors score and categorize customers. Customer behavior is scored. Customers are categorized into two groups:

- Unknown/Unconverted
- Converted

If a customer is in the positive samples segment, the customer is in the Converted group.

Based on the thresholds, adjusted by a seek bar located under the histogram, each of the customer profiles is assigned to one of four grades:

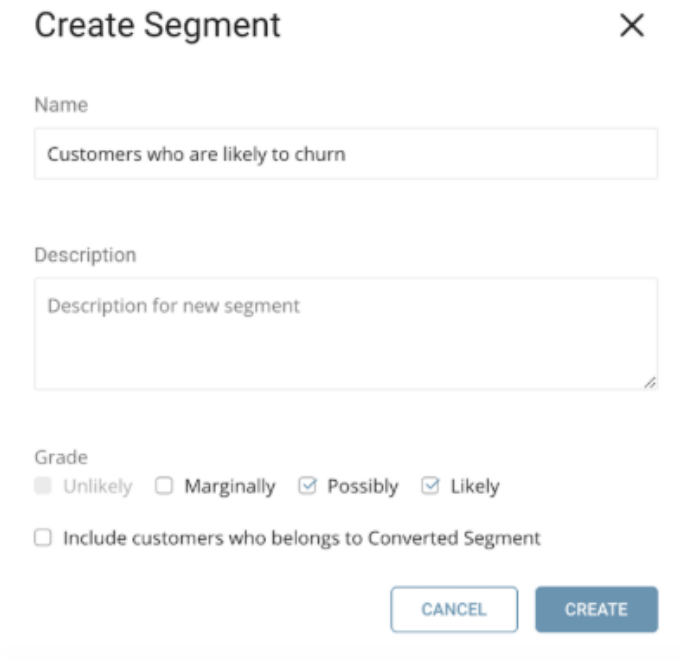
Likely	Possibly	Marginally	Unlikely



For example, no active customers are categorized into the Likely grade, and 29 active customers are in the Possibly grade. If you like to reach to more “likely” customers, you must adjust the right-most threshold to smaller value on the seek bar so that the percentage in the Likely circle is increased to a higher value.

Create a New Batch Segment Based on the Predictive Scores

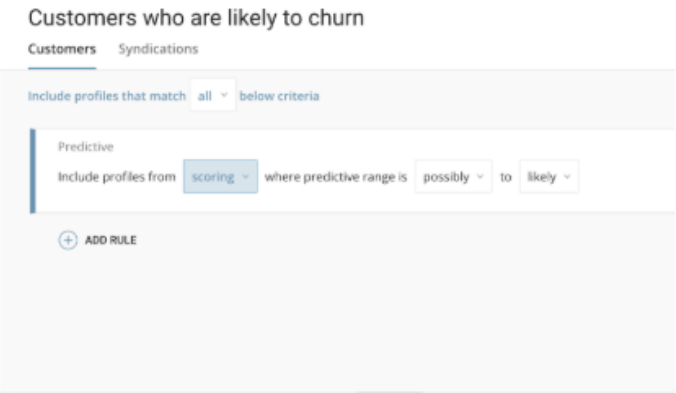
After thresholds are adjusted to desired positions, select Create New Segment:



You create a new batch segment based on the predictive scores. For example, you are interested in Possibly and Likely customers.

You can also specify if the segment includes customers who are in the positive sample, population segment, or in both.

A new segment based on the predictive scoring is created as follows:

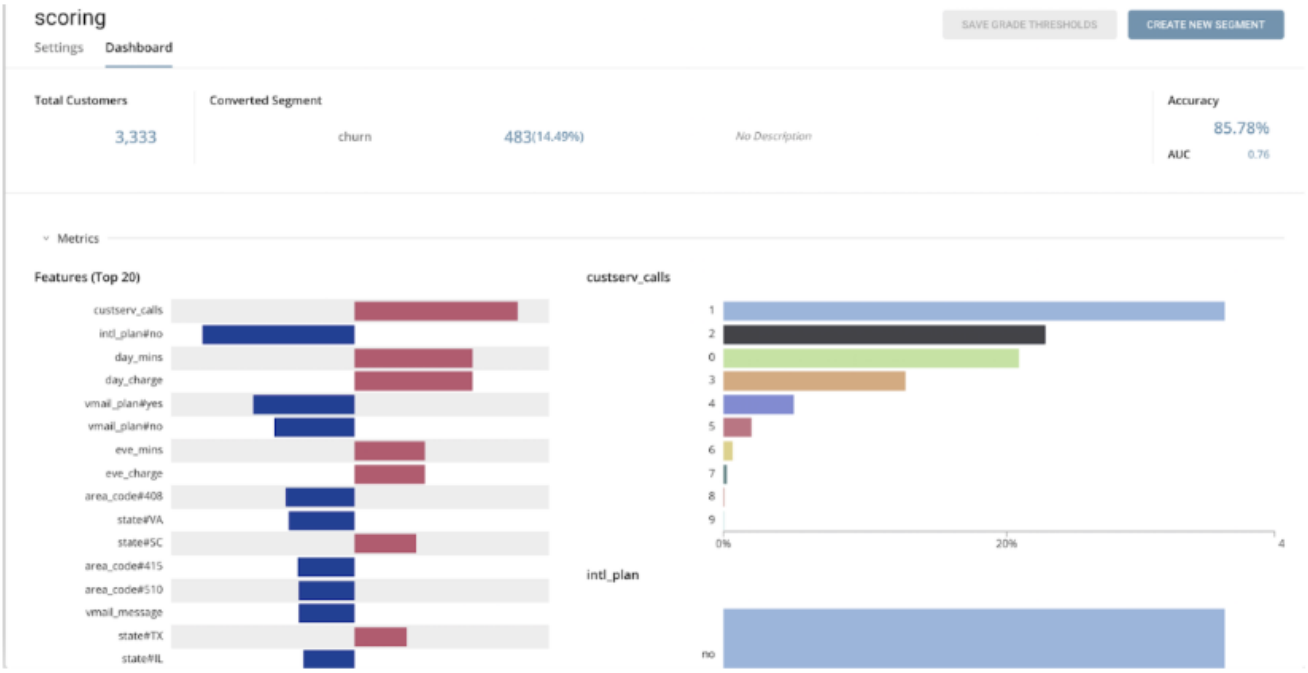


Because the Possibly and Likely grade respectively have 0 and 29 customers according to the dashboard, this batch segment contains 29 “promising” customers in total.

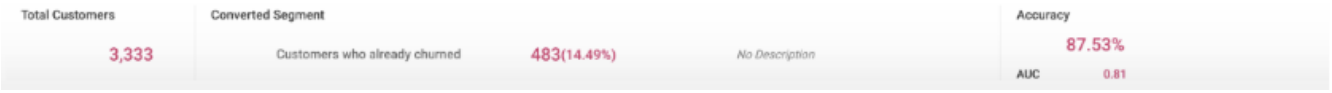
Set up activation on the segment. For example, you can send an email (such as special campaign information) to the customers who might churn in the near future to prevent their churn.

Understand Predictive Model and Tune It to Achieve Higher Accuracy

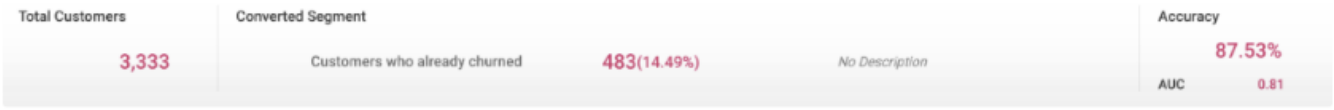
Machine learning on real-world data is not simple, and is sometimes inaccurate or results in an undesired prediction result. Use auxiliary information, provided at the bottom of the Predictive Scoring view to understand and improve your predictive scoring:



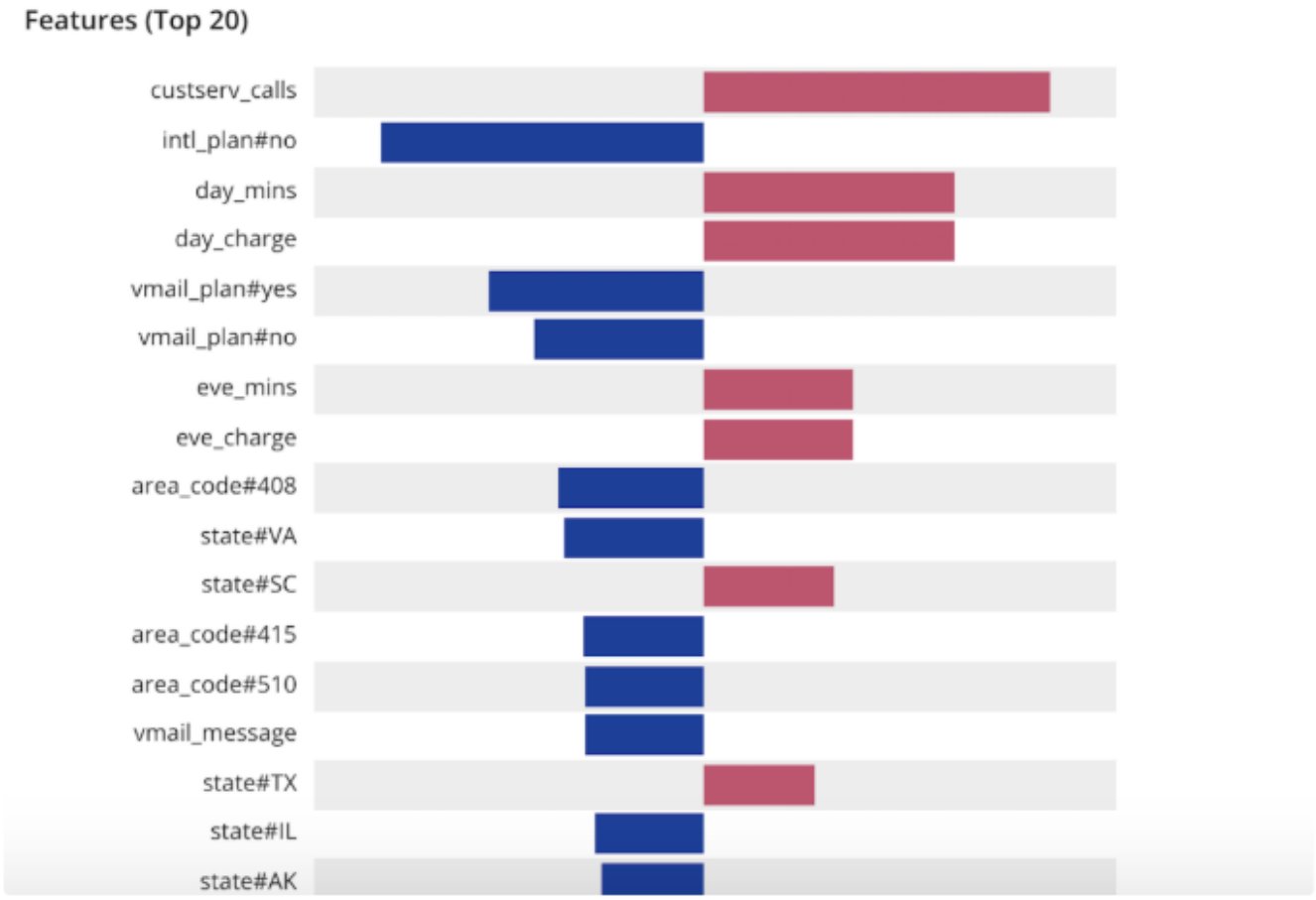
Statistics of your audience are shown with an estimated accuracy of prediction:

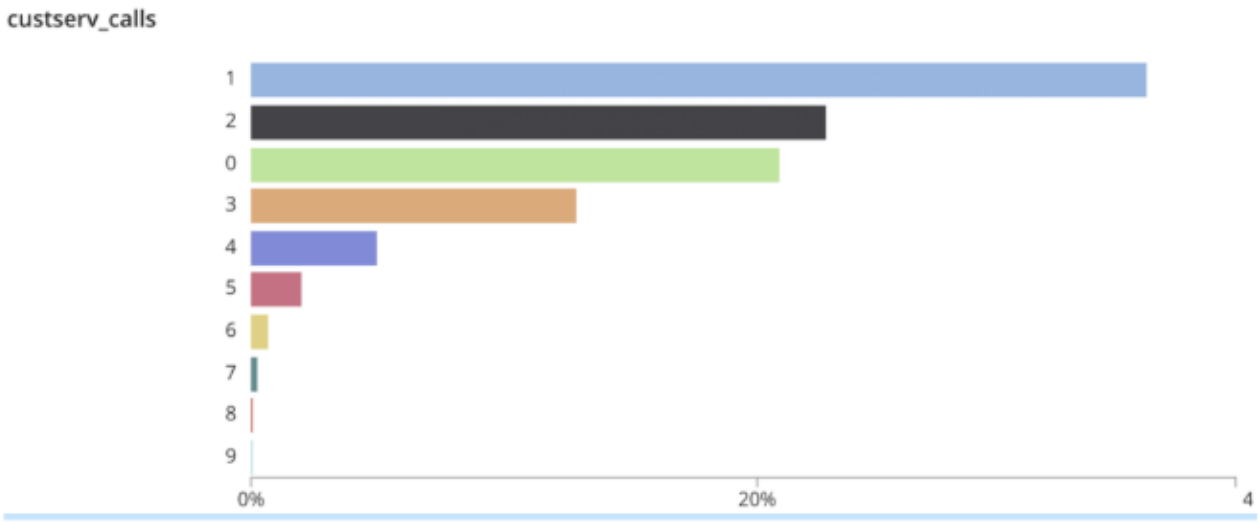


Statistics of your audience are shown with an estimated accuracy of prediction:



Plus, you can visually confirm which attributes strongly contribute to the prediction and what kind of values exist in each attribute:





Reviewing the information, you can see that **customer service calls** positively contributes to customer churn, and that **no international plan** leads to lower churn rate.