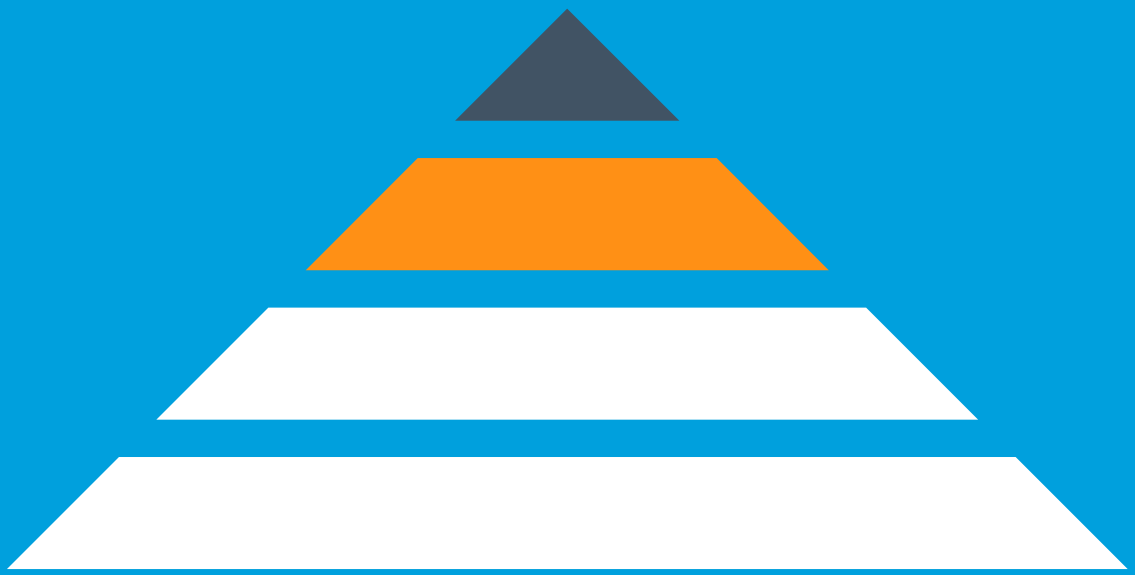
ORACLE® + DATASCIENCE.COM

An Executive's Guide to Predictive Data Modeling

An introductory look at how data modeling can drive better business decisions.



Introduction

Executives are making multi-million dollar decisions every day, but how many of those decisions are data driven? Less than you may think.

While business leaders seem to have reached a consensus about the importance of big data to business, only 29% of c-suite level executives are basing their decisions primarily on data analysis. And more than half (52%) admit to downplaying the importance of that analysis due to a lack of understanding.[1]

However, with big data driving a $65.7 million boost in net income for Fortune 1000 companies that increase data accessibility by just 10%, decision makers can no longer afford to ignore what they don't understand.[2] That's where data modeling comes in.

A good data model will bridge the gap between your analytics team and management. It allows for communication of the same information at different levels of detail, from data about a single customer's shopping habits to a comprehensive look at how the features on your site influence customer behavior and drive revenue growth.

So how do you start data modeling in a way that is meaningful to your business? This guide will provide an introductory look at how data modeling works, and demonstrate how the process can be used to overcome a real business challenge: preventing customer churn.

[1] PwC, "Guts & gigabytes," 2014. | [2] Baseline magazine, "Surprising Statistics About Big Data," February 2014.

What is Data Modeling?

Data modeling for business starts with a dataset and ends with a framework that can be used to understand the processes of your business mathematically – and to make predictions about how it will operate in the future.

The data modeling process is a technical one, and will likely be conducted by a data scientist or statistician. However, creating a predictive model that is truly representative of your business requires input from the people who understand your operations and business challenges – so it's imperative that you participate in the process.

To help you understand what that process is, here's a simplified step-by-step breakdown of how to create a data model (and how to begin using it to make business predictions):

Clean your data

Remove inaccurate or irrelevant information from your dataset. Incorrect data can lead to false conclusions when you're modeling.

Identify features

Extract relevant "features" from your dataset. This process, called feature engineering, involves combining your raw data into categories like customer spending per day or number of days since conversion, depending on what data you have and the scenarios you're ultimately trying to model. This process will make your data more palatable to your algorithm.

Select an algorithm

An algorithm is essentially a set of instructions for your computer system that tells it how to solve a mathematical problem. There are many algorithms to choose from; the type you select will depend on what you're trying to model, the data you have, and the computational power available.

Create and validate your model

Before you can start making predictions, you have to create a model and ensure that it works properly. You can do this by splitting your data into two different groups: training data and testing data. The training dataset will be used to "train" your algorithm – the result of this training process is your model – and the testing dataset can help you assess how well the model is able to make predictions.

Generate insights

Once your model is in working order, you can use it to understand your historical data (for instance, you can identify which types of customers have historically been the most valuable to your business) as well as make predictions about the future (how long those

Getting Started:

Modeling Customer Churn

Now that we've broken down the modeling process, it's time to dive into a real-world example. As you well know, every business wants to keep its customers coming back. Luckily, data modeling can help you understand when and why your customers defect from your service.

THE PROBLEM

You want to understand which traits are related to shorter subscriptions and pinpoint the behaviors that precede cancellations so you can do a better job of identifying customers who might leave your service – and reach out to them before they do.

> Your customers are canceling their subscriptions to your product.

Here's the data you might need to accomplish this:

Email Campaign History Data

The type of email (for example, promotions or discounts), when customers received the email, and recipients of each email campaign

User Demographic Data

Zip code, gender, age, income, education level, and other socio-economic characteristics

User Behavior Data

Date of subscription, login activity, transaction history, brand preferences, page views

THE APPROACH

One way to figure out who is going to defect from your service is through a survival analysis. A survival analysis uses various data models to predict the probability of an event – in this case, the probability a customer will keep using your service after a certain amount of time (a metric called "survival rate").

So, what determines a customer's survival rate? You're about to find out.

Step 1

First, let's identify some features:

Here are some examples of features you can extract from your raw data (these are all on a "per user" basis):

- Customer spending per day
- Customer spending per visit
- Zip code/county
- Gender
- Age
- Income
- Number of purchases per month
- Number of times the customer has accepted a promotion

Step 2

Now let's choose an algorithm:

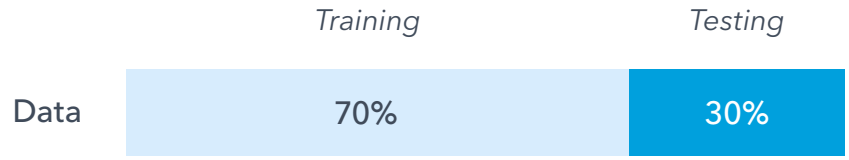
Remember, you'll be selecting an algorithm that can help determine a customer's survival rate. Cox's Proportional Hazard is a solid choice, because it measures the effects of different variables (like the features above) on survival.

Different types of algorithms will help you determine different things. For example, a regression algorithm works well for predicting time-driven events, while clustering algorithms are great for grouping customers who share similar characteristics (a process called customer segmentation).

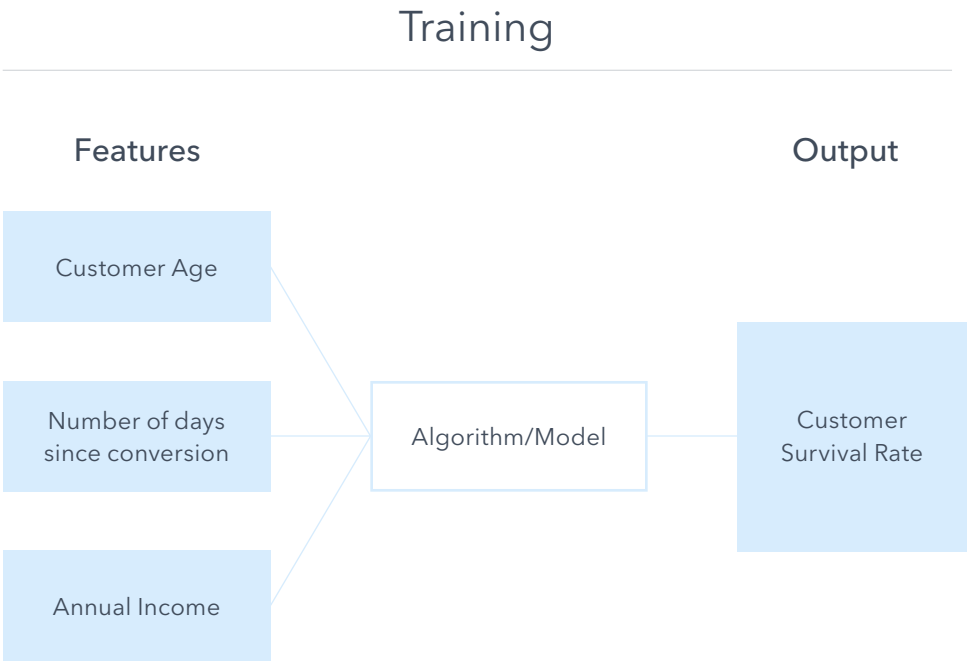
Step 3

Lastly, let's create and validate your model:

Your model is created when you train your algorithm. Before you start that process, you'll have to partition your data into two groups: a training dataset and a testing dataset.

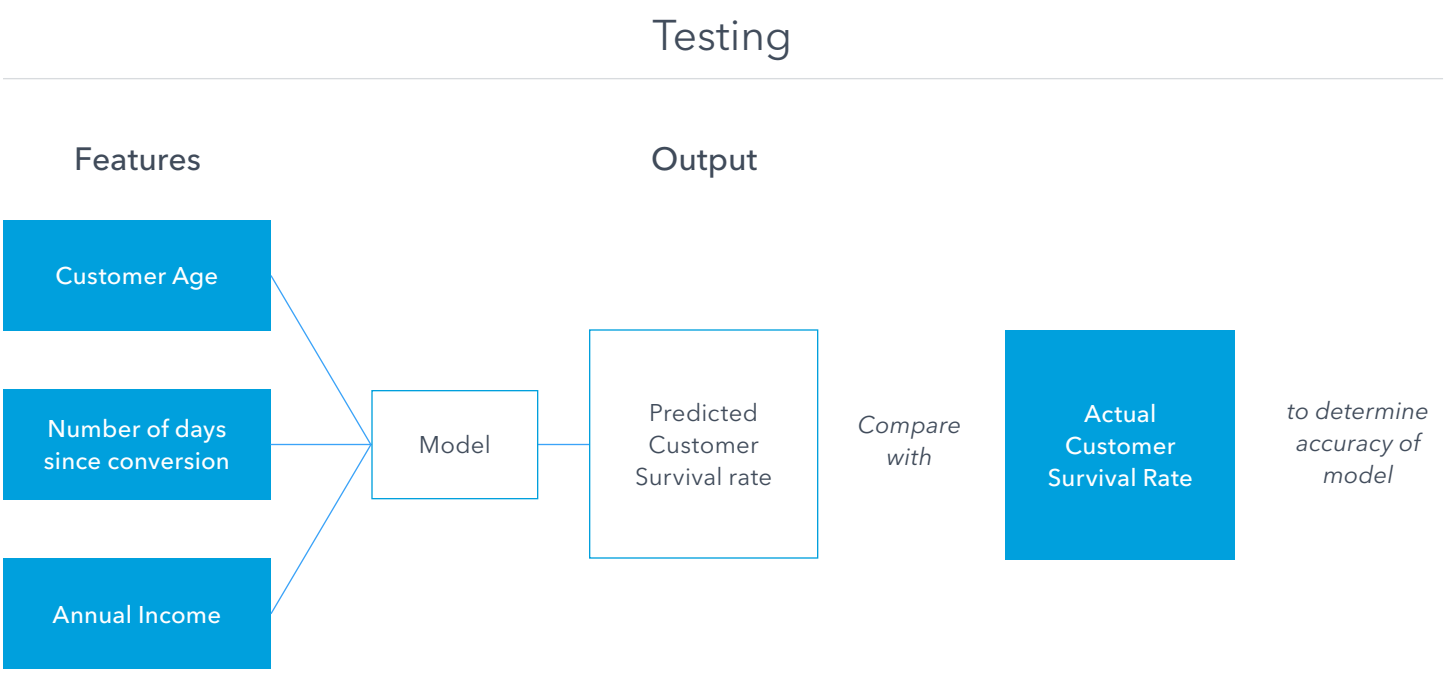


The training dataset, unsurprisingly, is used to train your algorithm (it's also larger than the testing dataset because you want to make your model as experienced as possible). This dataset contains all of the features (like customer age or days since conversion) related to your chosen business challenge – as well as your desired outcome (the survival rate of your customers). Using this data, the algorithm can learn how features come together to form the outcome you want. During this process, you will create your model.

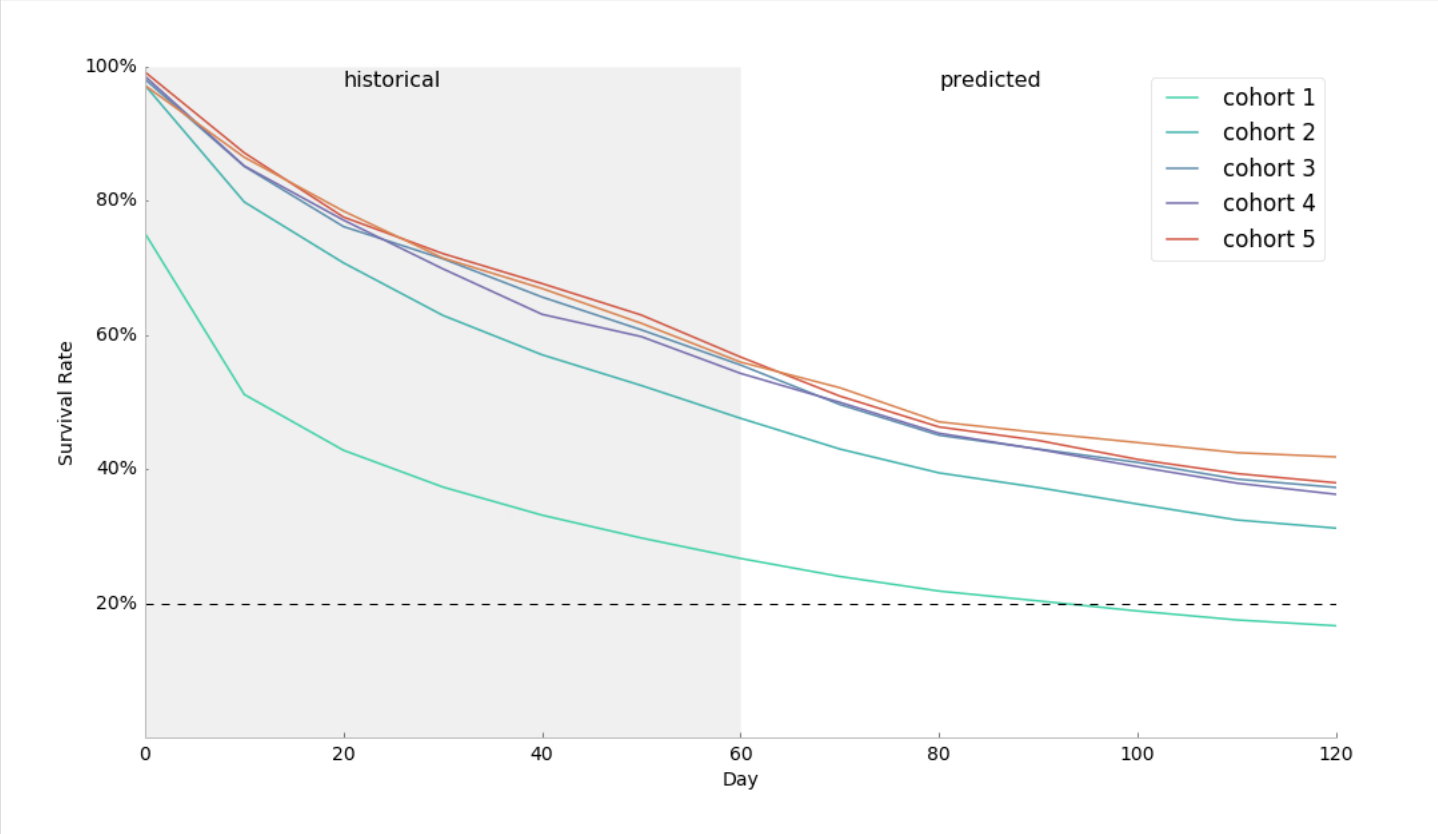


But the work doesn't stop there. Because you ultimately want a model that can make predictions about data it doesn't have, you'll have to test how accurately it can do just that. The testing dataset will also contain the features you've already identified and data about the survival rate of your customers – but you will only provide the model with feature data.

When you run the model with the testing dataset, the output will be a prediction about your customers' survival rates. And because you already have the actual data about those survival rates (you just didn't give them to the model!) you will now know how accurate your model is. This process is called cross validation.



The Outcome



In this chart, the gray region indicates the historical data that was used to train the model, while the white region is the predicted data from your model.

Once you've ensured that your data model works, you can begin to forecast the survival rate of every user of your service for the next month and beyond. In the chart above, you can see that of the first cohort of customers, only 20% will be returning to your service 95 days after they first subscribed. Once a cohort has reached this threshold, it might make sense for you to send a targeted email campaign designed to bring them back to your service.

As a byproduct of this analysis, you'll also understand which traits define your most valuable users, allowing you to identify (and reach out to) potential customers with attributes that make them more likely to spend money with your business for longer.

Conclusion

If you have the right data, modeling will take much of the guesswork out of your business decisions, no matter what industry you serve. From financial services to manufacturing, predictive analytics now provide a competitive advantage for businesses looking to boost revenue or lower costs.

Predictive modeling can be applied to any number of scenarios, including cart abandonment, search optimization, lifetime value prediction, or fraud detection.

To find out how Oracle's DataScience.com platform can help your business build and deploy predictive models, visit **www.datascience.com**.

ORACLE + DATASCIENCE.COM

Connect with us on social media:

@DataScienceInc

https://www.linkedin.com/company/datascienceinc

@DataScienceInc

https://www.facebook.com/datascience

Oracle Corporation, World Headquarters 500 Oracle Parkway Redwood Shores, CA 94065 USA