ŷhat

Data Science in Practice

Five common applications of data science with concrete, real-life use cases

Presented by
Yhat
http://yhat.com/
June 2016

What is data science...actually?
How do real companies use data science to make products and operations better?
What does the data science lifecycle look like?

In our first whitepaper, “Applied Data Science,” we translated the hype-y term of data science into the plain english definition of “using data to make better decisions, optimize processes and improve products and services.” We also described the central goal of data science: getting statistical models into production.

In this whitepaper we introduce five common applications of data science that build upon those concepts. Our goal is to debunk the impression that data science is some type of obscure black magic and give you concrete examples of how it is applied in reality. You’ll learn how real companies are using data science to make their products and day-to-day operations better. Last but not least, we describe the data science life cycle and explain Yhat’s role in getting models into production.

APPLICATION 1: RECOMMENDER SYSTEMS

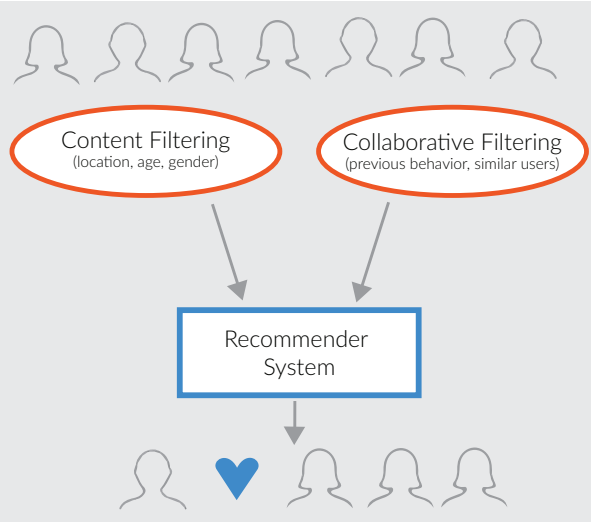
Recommender systems, also known as recommender engines, are one of the most well known applications of data science. Recommender systems are a subclass of information filtering systems, systems that cut through the noise of all options and present users with just the subset of options they'll find appealing. The data being filtered can range from products on an e-commerce site to dating matches that appear as you search for 'the one.'

WHAT IS A RECOMMENDER SYSTEM?
A model that filters information to present users with a curated subset of options they're likely to find appealing

HOW DOES IT WORK?
Generally via a collaborative approach (considering user's previous behavior) or content-based approach (based on discrete assigned characteristics)

WHAT IS A REAL USE CASE?
Tendril uses recommendation models to match eligible customers with new or existing energy products

Recommender Systems



Recommender systems offer a more intelligent approach to information filtering than a simple search algorithm by introducing users to items they might not have otherwise discovered. Recommender systems generally take either a collaborative or content-based approach to filtering. Collaborative filtering considers a user's previous behavior, as well as the behavior of similar users. Content-based filtering provides recommendations based on discrete attributes or assigned characteristics.

Data scientists at energy software company Tendril opted for a hybrid approach that combines both collaborative and content-based filtering. Tendril provides analytics and consumer solutions to energy suppliers, including which energy products consumers would most likely consider. "We use Support Vector Regression models to predict household energy consumption to provide our clients with in-depth, personalized information about their customers," explains Mark Gately, Data Analytics Manager at Tendril. "This detailed information is also used in recommendation models, which help match eligible customers with new or existing energy products."

APPLICATION 2: CREDIT SCORING

If you have ever applied for a credit card or a loan, you're likely already familiar with the concept of credit scoring. What you may be less aware of is the set of decision management rules evaluating how likely an applicant is to repay debts behind the scenes.

WHAT IS CREDIT SCORING?
A model that determines an applicant's creditworthiness for a mortgage, loan or credit card

HOW DOES IT WORK?
A set of decision management rules evaluates how likely an applicant is to repay debts

WHAT IS A REAL USE CASE?
Ferratum Bank uses machine learning models to reach prospective customers that may have been overlooked by traditional banking institutions

The first general purpose credit scoring algorithm, now known as the FICO score, was introduced in 1989. The FICO score is still one of the most widely used models in the United States today, though peer-to-peer and direct lending organizations have focused on developing new techniques over the past few years. These new machine learning models and algorithms capture innovative factors and relationships that traditional loan scorecards couldn't, like how applicants manage monthly cash flow or whether friends or community members would endorse the applicant.

One such company is Ferratum Bank, a pioneer in financial technology and mobile consumer lending since 2005. "We developed complex statistical and machine learning models to enable smarter lending decisions," explains Scott Donnelly, Director of Business Lending at Ferratum Bank. "By getting creative with our approach and adopting innovative technologies, we've been able to reinvent how both consumers and businesses obtain loans. This has allowed us to reach prospective customers that in the past may have been overlooked by traditional banking institutions."

APPLICATION 3: DYNAMIC PRICING

You walk out of the store, arms full of groceries, only to realize that a torrential downpour began as you perused the produce inside. You struggle to retrieve your phone, check your favorite ride app and are dismayed to find...a 2.1x surge!? Welcome to your first lesson on dynamic pricing.

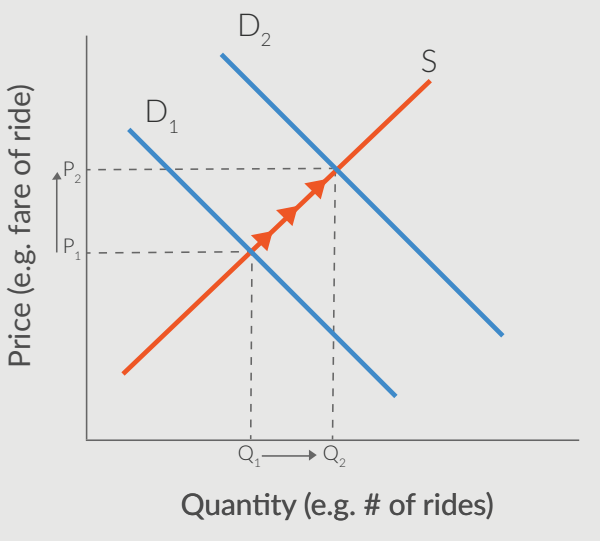
WHAT IS DYNAMIC PRICING?
Modeling price as a function of supply, demand, competitor pricing and exogenous factors

HOW DOES IT WORK?
Generalized linear models and classification trees are popular techniques for estimating the "right" price to maximize expected revenue

WHAT IS A REAL USE CASE?
Turo uses dynamic pricing models to suggest prices to the people who list and rent out cars

Businesses use dynamic pricing algorithms to model rates as a function of supply, demand, competitor pricing, and exogenous factors (e.g. weather or time). Many fields, from airline travel to athletics admission ticketing, employ dynamic pricing to maximize expected revenue. The nuts and bolts of dynamic pricing strategies vary widely, though generalized linear models and classification trees are popular techniques for estimating the "right" (lowest/highest) price that consumers are willing to pay for a book, a flight, or a cab.

Dynamic Pricing



Turo, a peer-to-peer car rental service operating in over 2,500 cities, uses dynamic pricing to suggest prices to the people who list and rent out their cars on the platform. "Dynamic pricing helps us to balance supply and demand and ensure that both our travelers and our hosts are getting a fair market deal," explains Jérôme Selles, Director of Data Science and Analytics. "Three years ago we started to model supply and demand dynamics, so working on dynamic pricing was an intuitive next step."

"We quickly realized that the gap between model development and model deployment, in production, was much bigger than expected. It requires a very wide spectrum of skills: from knowledge of statistical modeling to software architecture best practices. We use Yhat's platform, ScienceOps, to transform our dynamic pricing prototype into a production-ready algorithm in the languages our Data Science team prefers to work in, R and Python."

APPLICATION 4: CUSTOMER CHURN

Churn rate describes the rate at which customers abandon a product or service. Understanding customers' likelihood to churn is particularly important for subscription-based models, everything ranging from traditional cable or gym memberships to recently popularized monthly subscription boxes.

Data scientists looking to predict customer churn may consider a variety of algorithms for the job, such as support vector machines, random forest, or k-nearest-neighbors. Beyond the accuracy of a given model, data scientists must also balance the tradeoff between precision (correctly predicting a

churning customer) and recall (how many predictions were actually successful). So what's better? Classifying every churning customer but occasionally mislabeling a non-churning customer? Or identifying fewer churning customers, but not mislabeling non-churners? It's a difficult decision that requires in-depth knowledge of the business case and years of experience.

WHAT IS CUSTOMER CHURN?
Predicting which customers are going to abandon a product or service

HOW DOES IT WORK?
Data scientists may consider using support vector machines, random forest or k-nearest-neighbors algorithms

WHAT IS A REAL USE CASE?
EAB combines data from transcripts, standardized test scores, demographics and more to identify students at risk of not graduating

These are familiar questions for the data scientists at EAB, the education division of The Advisory Board Company. EAB provides data driven applications and insights to hundreds of institutions of higher education. "A key component of our Student Success Collaborative product, used by academic advisors and other administrators, is a predictive model of student graduation," says Harlan Harris, Director of Data Science. "We combine data from transcripts, standardized test scores, demographics, and other facts about students to provide a graduation risk score. Colleges and universities use these scores to identify students at risk of not graduating more efficiently, so they can intervene and help those students graduate."

APPLICATION 5: FRAUD DETECTION

Financial technology, or 'FinTech,' companies offer financial services like banking, investing, and payment processing via software, rather than through traditional banking institutions. Companies processing massive volumes of financial transactions also need a quantifiable way to detect and prevent fraudulent transactions from being processed.

WHAT IS FRAUD DETECTION?
Detecting and preventing fraudulent financial transactions from being processed

HOW DOES IT WORK?
Fraud detection is a binary classification problem: "is this transaction legitimate or not?"

WHAT IS A REAL USE CASE?
Via SMS Group uses a combination of complex data lookups and decision algorithms written in R and implemented in PHP to assess whether a loan applicant is fraudulent

Traditional fraud detection presents a fairly straightforward problem: Is a transaction legitimate or not? Otherwise called a binary classification problem. This can be trickier than it seems, especially when you have thousands (or even millions) of legitimate transactions occurring for every instance of fraud. To add insult to injury, a single occurrence of fraud can cost a company an exorbitant amount of money. To combat this,

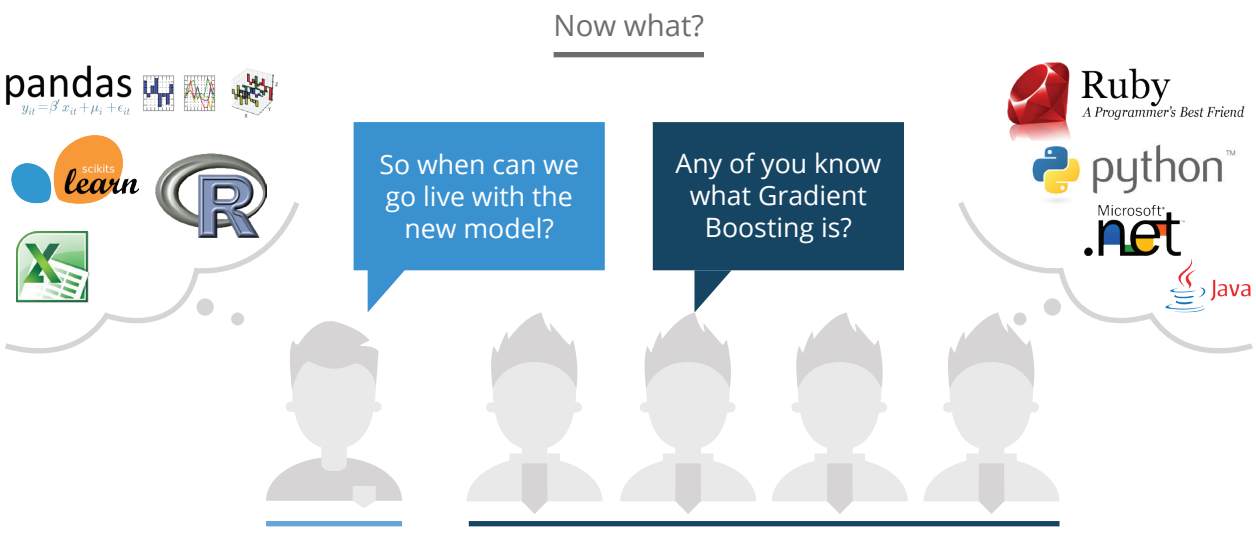
some data science teams pair supervised classification techniques with anomaly detection algorithms to identify outliers and pick out suspicious behavior.

Dmitrijs Ļvovs is responsible for managing risk at VIA SMS Group, where over €60MM is loaned to consumers across 6 countries every year. The risk analytics team at VIA SMS Group use advanced algorithms to assess whether an applicant is fraudulent prior to considering whether or not to underwrite the requested loan. "We write our decision algorithms in the R programming language and implement them into our web and mobile apps in the server-side language of PHP. By using R, we can leverage a combination of complex data lookups and state of the art algorithms to identify fraudulent transactions," explains Dmitrijs.

THE DATA SCIENCE LIFECYCLE

As the five applications above demonstrate, data can be used to inform decisions, optimize processes, and improve products and services across a very wide range of business problems. Regardless of the specific question at hand, the data science lifecycle always culminates with selecting the winning model strategy and implementing it into an application where real business value can be realized.

For data scientists who build their models in open source programming languages like R and Python, the path to production application is not always clear. Mobile and web applications are built using platforms and frameworks like .NET, Ruby on Rails, Java, PHP or Node.js, which cannot consume models written in R or Python. As a result, many models are abandoned after months of work before they ever see the light of day. Alternatively, data scientists' advanced statistical procedures may be "tossed over the fence" to engineers and manually recoded into another language, a notably



difficult, time consuming and error-prone process.

YHAT'S ROLE

Yhat's data science operations system, ScienceOps, eliminates the counterproductive barrier between data scientists and engineers by making R and Python models accessible via REST API. Instead of translating models, data scientists can deploy models to ScienceOps, where automatically generated API endpoints make production integration quick & easy.

WHAT IS SCIENCEOPS?
Yhat's data science operations system that eliminates the barrier between data scientists and engineers

HOW DOES IT WORK?
ScienceOps makes R and Python models accessible via REST API and provides a platform to monitor, manage and scale data science models

WHAT IS A REAL USE CASE?
ScienceOps is used by companies around the globe, including each of those highlighted in the five applications above

Our core mission at Yhat is to allow data scientists to deploy predictive models rapidly, frequently and reliably, but we recognize that a data scientist's job does not end there. Beyond the initial step of implementing models, ScienceOps also provides the ability to monitor, manage and scale models.

Companies around the globe, including each of those highlighted in the use cases above, rely on ScienceOps to take data science models from prototype to production.

To find out more about how your business can deploy models rapidly, frequently and reliably with ScienceOps, get in touch with the Yhat team or schedule a demo today.

Works Cited

Chiang, Eric. “Predicting Customer Churn with Scikit-learn.” The Yhat Blog. Yhat, 20 Mar. 2014. Web.

Huang, Cheng-Lung, Mu-Chen Chen, and Chieh-Jen Wang. “Credit Scoring with a Data Mining Approach Based on Support Vector Machines.” Expert Systems with Application 33 (2007): 847-56. Web.

Leskovec, Jure, and Jeffrey Ullman. “Recommendation Systems.” Mining of Massive Data Sets. Ed. Anand Rajaraman. 2.1 ed. Cambridge: Cambridge UP, 2014. 307-41. Print.

Phua, Clifton, Vincent Lee, Kate Smith, and Ross Gayler. “A Comprehensive Survey of Data Mining-based Fraud Detection Research.” Web. <https://arxiv.org/pdf/1009.6119.pdf>.

Yhat. Applied Data Science: Practical Guide to Building Data-driven Products beyond Analysts’ Laptops. New York: Yhat, 2014. Print.

About

Yhat (pronounced Y-hat) provides an end-to-end data science platform for developing, deploying, and managing real-time decision APIs.

Yhat's flagship product, ScienceOps, enables data scientists to transform static insights into production-ready decision making APIs that integrate seamlessly with any customer- or employee-facing app. Yhat also created Rodeo, an open source integrated development environment (IDE) for Python.

Contact Us

http://yhat.com
info@yhathq.com

(718) 855-2107

45 Main Street
Suite #707
Brooklyn, New York 11201