Nowadays, companies of all sizes rely heavily on data science and business analytics to solve their business problems.

While data science has worked wonders for many companies, the same can't be said for others. When applied to any business, data science can be a double-edged sword. If not implemented correctly, data science can lead to mistakes that cost your business thousands if not millions.

What's the ROI on Data Science and Machine Learning?

Before we can start addressing the ways in which data science can ruin your business, let's see how it can help it. After all, despite the possible risks, the ROI on data science and machine learning can be quite high (which is why most companies are implementing it).

Let's consider two companies (Amazon and Google AdWords) that managed to reap the benefits of integrating data science in business.

Amazon

In 2017, Amazon had a total revenue of $177 billion. Many insiders would argue that around 20% (roughly $35 billion) of their revenue that year was generated as a result of their cross-selling and recommendation efforts, which are based on data science.

AdWords

In 2017, Google AdWords' total revenue was $95 billion. This is actually a bigger success story than Amazon. How so? Almost 100% of that revenue was generated through machine learning and business analytics.

When you use your browser to search for something, Google displays an advertisement that matches your interests, making you more likely to click on – and purchase – the advertised product.

However, there are cases in which data science in business can backfire. Like in the case of Tesco.

Tesco

As early as the 1990s, Tesco began using prediction models and big data analytics to improve their advertising efforts. The British company was able to provide better, more personalized ads, which impressively grew their profits more than 7x within two decades.

Several years ago, however, things started to go south. Tesco's customers felt like the amount of data they had to share continued to grow, while the return on value for them was low. In a nutshell, the predictive models Tesco created, ultimately turned their own customers against them.

Tesco's failure can't only be attributed to their machine learning models as they also failed to enter the US market. In the end, this example shows that businesses should consider how the models they are creating are utilized with regard to creating value for their customers.

These 3 examples show how machine learning models and business analytics can boost or hurt a business.

3 main problems when applying machine learning in business

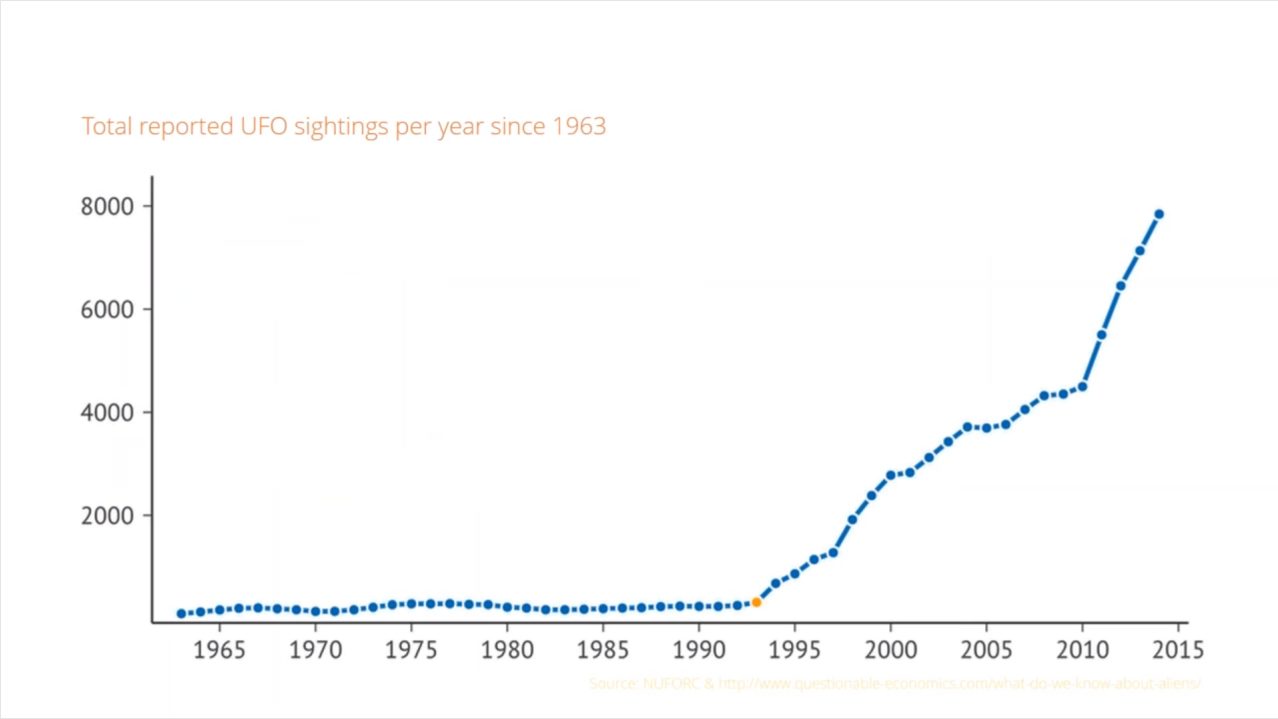
There are several mistakes that you should avoid when applying data science in business but we will focus on three specifically.

1. **Confusing correlation and causation:** Highly-correlated features can overshadow the patterns your machine learning model is supposed to find. This can lead to models which perform worse in production than during the model building.
2. **Incorrectly validating a model:** This can lead to over-optimistic estimations of your model accuracy. Even if you model your data correctly, validating your model in the wrong way means you won't understand how well your model will work in production.
3. **Focusing on models, not data:** Creating complex models instead of thinking about the data and the problem you are trying to solve can be inefficient, inaccurate, and hard to explain. Complex models are also less robust against data from a changing world.

Business Mistake # 1: Confusing Correlation and Causation

What UFO sightings tell us about the dangers of Data Science

The chart below shows the total number of UFO sightings per year since 1963. The data was taken from the database of the National UFO Reporting Center and this work was originally done by work by Dan Henebery and Josiah Davis.



Do you notice anything unusual? The number of sightings was steady for several years and then increased dramatically in 1993.

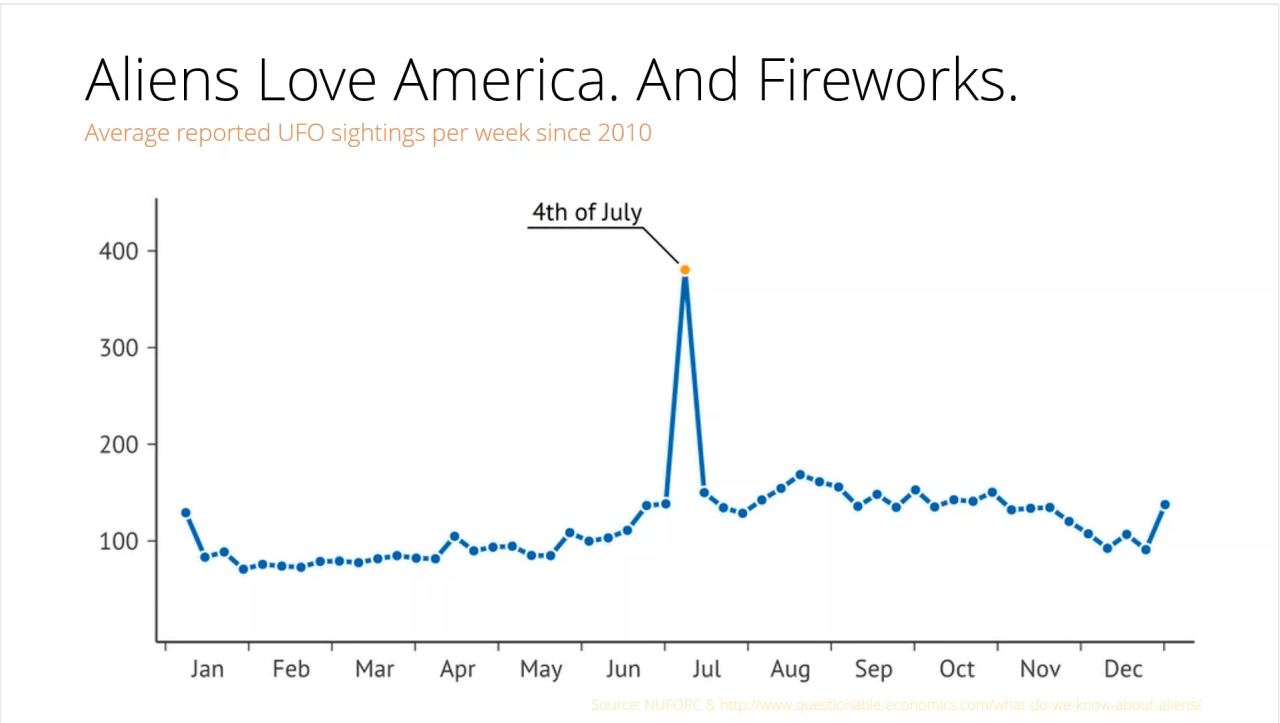
In September 1993, the first episode of *The X-Files* aired. At its peak, more than 25 million people in the US watched it. So what can we deduce from this chart? Without any additional information, the most likely explanation has to be that aliens were big fans of *The X-Files* and came to Earth to watch the series with us.

Next let's take a look at the data below that shows the frequency of UFO sightings based on the time and day of the week (Monday through Sunday). Yellow-orange represents more frequent sightings.

You probably notice that most UFO sightings happen on Saturday nights. And it can't be a coincidence that most parties happen on Saturday nights as well.

We already knew that they like to watch *The X-Files*, but now we also know they like to party with us too.

Next, let's look at the average number of reported UFO sightings per week since 2010.



Most sightings happen during the week of 4th of July. Perhaps aliens love America and enjoy fireworks as well.

At this point, you've most likely realized that we're being ridiculous. We can't jump to these conclusions by just considering this specific data. This is a demonstration of how correlation and causation can easily be confused. It is easy to spot this mistake here but much harder for most business relevant scenarios.

Another example of mistake #1

In 2001, a study by Neuman & Celano showed that children who had access to more books received better grades at school.

Supported by a $20 million gift from the William Penn Foundation, Philadelphia began a five-year project to transform 32 neighborhood branch libraries. The project's goal was to "level the playing field" for all children and families in Philadelphia.

At first glance, the plan looked great. Unfortunately, the study that it was based on did not take into consideration whether the children actually read the books. The only thing that was considered in the study was whether the books were available or not.

That's a typical example of how easy it is to take a correlation and draw the wrong conclusion, and shows how data science in business can turn out to be detrimental. It this case, it led to a spending of $20M which could have been better used in a different way.

How to Avoid Confusing Correlation and Causation

1. Check for correlations

Data science requires human thought, so keep the human in the loop. Don't just trust a ready set-up model to do everything you request. Always include human thought to truly understand where and how a specific model will be used. This means that you should check for correlations in the data and understand what they mean before modeling. You should also consider removing factors with too high correlations to the target or to other factors.

2. Take out information that is only available after the point of prediction

The best models are not the ones that give you highest correlation, but enough information early enough to act on. Some data can be highly correlated because it's available after the fact, meaning it's too late to act on it. In such cases, it's better to remove that data from the model, despite that making the model less precise.

There is no point in making the model look perfect during the building period but then fail miserably in production. It is better to have a realistic estimation of how well the model will perform later on instead of getting a negative surprise.

Mistake # 2: Incorrectly Validating a Model

When it comes to model validation, the topics get a lot more technical, and it's much easier to get things wrong.

A Customer Churn Example

Imagine a company that loses $200 million per year due to customer churn. They create a machine learning model that they believe can reduce churn by 20%, saving them $40 million. However, to save those customers from churning, the company needs to spend $20 million to take the right measures to retain them. So, the company makes the call that it's worth it to spend that $20 million to save $40 million from churn.

Once this company put the model into production though, they realized the model was over-optimistic and it only reduced customer churn by 5%. So, they were only able to save $10 million but spent $20 million doing so. The consequences of incorrect model validation may not always be a loss of $10 million but models performing worse in production is a frequent issue.

It is important to notice that the model did actually not change its behavior here. It is also not a simple case of overfitting. All what happened is that the model always performed at the lower accuracy level, but due to a wrong validation, it looked much better during model building. The truth was only discovered after the model was put in production which is of course extremely risky.

How to Avoid Incorrect Model Validation

Here we summarize how to avoid this but we also have a whitepaper explaining correct model validation in more detail.

1. Ignore training errors

Training errors can be misleading because they are calculated by the same data used in training the model, so they deliver an overly optimistic estimation of model accuracy. Instead, use a test error, using two completely disjoint data sets, to estimate how well the model will perform for new and unseen cases in the future.

2. Prefer cross-validation whenever possible

A k-fold cross-validation is the go-to method whenever you want to validate the future accuracy of a predictive model. It is a simple method which guarantees you that there is no overlap between the training and test sets. It also guarantees that there is no overlap between the k test sets which is good since it does not introduce any form of negative selection bias.

And last but not least, the fact that you get multiple test errors for different test sets allows you to build an average and standard deviation for these test errors. This means that instead of getting a test error like 15% you will end up with an error average like 14.5% +/- 2% giving you a better idea about the range the actual model accuracy will likely be in when put into production.

It is ok to use a single hold-out set in the early stages of the prototyping phase for runtime reasons. However, before you put the model into production you should consider a full cross-validation whenever possible.

3. All data transformations across rows must be inside of the cross-validation

You can contaminate your training data set by applying data transformations before the cross-validation which leads to information leakage about the test data into the complete data set. The only way to avoid surprises when going into production is to properly validate the model AND all data preparation which requires the data transformations to be part of the validation process. This also means that it is never a good idea to separate data engineering and model building between separate teams or tools.

4. Take out information that is only available after the point of prediction

This is a repeat from mistake #1 but think about it again during validation. If a model performs surprisingly well, this is often a sure hint that information was leaked. This can happen due to data transformations (see previous point) or by including highly correlated features which actually are information only available after the fact.

Mistake # 3: Focusing on Fancy Models Instead of Data

Using fancy models when they are not actually needed is a common mistake. There's a lot of hype revolving around data science and machine learning, especially when it's applied to business. As a result, many data scientists fall into the trap of focusing on creating complex models instead of preparing the data.

In business, it's crucial to use data preparation to improve your business processes. It doesn't matter whether you use a complex model or a simple one, as long as it saves your business money or increases profits.

Complex models are not per se a bad thing. Sometimes they are only model type which is powerful enough to create accurate predictions. But the complexity makes them harder to understand, which often renders them useless for business applications. They are often also not very robust against changes in the world. While they may perform accurately on the current data, the world is moving on and processes change all the time. Complex models often need more maintenance and retraining to keep up with the world while simpler models are performing more robustly.

Training time is also longer for complex models, which makes them less manageable for feature engineering and parameter optimizations. In order to establish some benchmark and understanding of the underlying processes, we recommend beginning with simpler models first and establishing some benchmarks. If understandability does not matter much for your applications, you can still move to more complex model types from there. What are the best practices for the model building process and making it work for your business?

1. Start with a plan and a question in mind

Take a step back and think about what you are trying to achieve with the data. What problem are you trying to solve? What are you trying to predict?

2. Define success before you start

We're used to defining KPIs for measuring success in business, but when it comes to machine learning, we often fail to define success metrics.

Success metrics are important because they allow you to benchmark models and think about potential improvements. They also help to detect potential mistakes with the model (e.g. if the model is overly-accurate, you know something isn't right).

3. Understand the problem you want to solve with machine learning

For this, you need to clearly understand your business problem, requiring close collaboration with your stakeholders.

4. Use common sense and invent new features

When you take a look at the data and understand the business problem, you should be able to create new features that can reveal a more accurate prediction. For example, say you are given a data set about whether your company will buy a lot of land. This data set has three attributes and one column that you are meant to predict – if the company bought the land or not. The three attributes are length, width, and price. By looking at the data yourself, none of these attributes seem to predict whether or not the company bought the land, so knowing what you do about the business you decide to calculate the area of the land, which is length*width. Then using that calculation, generate the price per square foot or meter.

5. Start with simple models

Simple models like linear regression or decision trees can be very powerful tools when analyzing data. In fact, they can sometimes outperform more complex models. Also, the more complex a model, the harder the results of a prediction are to explain. This makes your model look like a black-box to your stakeholders.

If your stakeholder or business partners aren't able to understand why and how such predictions occur, it'll be harder for them to trust you and your models.

How can you avoid these Data Science Mistakes?

Data scientists have started to use data science platforms to help them design, build, and visualize models that are easily understandable by a wide range of business professionals. Using a visual workflow designer for prototyping and validating predictive models can help you avoid the main mistakes that can occur when applying data science in business.

RapidMiner is a data science platform that truly aims to resolve the discrepancies between the world of data and business. Get started with a 30-day FREE trial.

And if you want to see more of it in action, check out this video on How to Ruin Your Business with Data Science and Machine Learning. We'd recommend skipping directly to the practical examples throughout the video.

In this post, we covered some of the most important data science problems that can ruin your business. However, more can occur when working with business data analytics on a daily basis. If you feel like we might have missed something, or you have any questions, let us know in the comments section below.