MODERN DATA SCIENCE: BEST PRACTICES FOR PREDICTIVE ANALYTICS

PeerPaper Report

10 TIPS FROM REAL USERS OF IBM SPSS MODELER

IT Central Station
Unbiased reviews from the tech community

ABSTRACT

Data science and machine learning provides the basis for business growth, cost and risk reduction and even new business model creation. Implementing predictive analytics does present some challenges, however. The process can be complex, and it can be difficult to find data scientists and analysts with a mix of the right skillsets. A drag and drop, visual data science tool, exemplified by IBM SPSS Modeler, enables rapid creation of machine learning models while making it easy to collaborate with data science and analytics teams as a whole. In this paper, members of IT Central Station who use IBM SPSS Modeler share their experiences and offer insights and recommended best practices for data science and machine learning.

CONTENTS

INTRODUCTION

Data science and machine learning provide the basis for business growth, cost and risk reduction and even new business model creation. Implementing predictive analytics does present some challenges, however. The process can be complex, and it can be difficult to find data scientists and analysts with a mix of the right skillsets.

A drag and drop, visual data science tool, exemplified by IBM SPSS Modeler, enables rapid creation of machine learning models while making it easy to collaborate with data science and analytics teams as a whole. In particular, IBM SPSS Modeler extends to the open source environment for data scientists who code in R and Python, where new innovation and custom algorithms can be built. In this paper, members of IT Central Station who use IBM SPSS Modeler share their experiences and offer insights and recommended best practices for data science and machine learning.

Data Science and Machine Learning Overview

The term "data science" refers to a collection of practices that leverage computer power to extract knowledge or insights from data. Businesses can harness predictive analytics, based on data science, to model behavior based on patterns. Done right, data science delivers value to businesses by enabling them to improve their understanding of operations, sales growth, customer experience and more.

For example, on IT Central Station, a Quantitative Researcher at a financial services firm with more than 10,000 employees described how his company used predictive analytics to estimate the lifetime value of customers. With data science, they can determine optimal approaches to customer acquisition, retention, cross-sell and up-sell as well as segmentation.

Other examples of data science benefiting businesses include:

- **Sentiment analysis**—Analyzing unstructured data in social media threads and product reviews to improve product messaging and merchandise selection.
- **Purchase intent**—Predicting who will buy a specific product and when the purchase will occur, based on past purchase behavior, browsing behavior, sentiment analysis, demographics and so forth.
- **Fraud detection**—Inspecting transactions and related data, like IP addresses of user devices, to

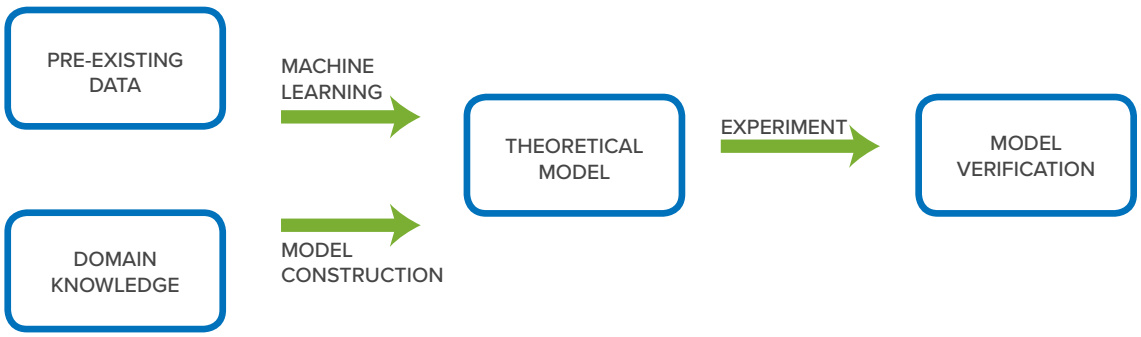

Figure 1 - High level flow of the predictive analytics modeling process.

determine if fraud or other improper activities are taking place.

- **Predictive maintenance in industrial operations**—Using data on past repairs and part replacements to predict when a part will wear out—and replacing it before there's a breakdown in operations or an accident.

Challenges to Data Preparation, Model Development and Training, and Deployment

A successful predictive analytics project doesn't just happen. It's the result of a series of process steps, each of which can be difficult and time-consuming. These include data preparation and development of the predictive analytics model, followed by the "training" of the model. The data in its raw form may not be useable. Without effective data preparation, model development and training, the predictive analytics may not work at all.

A number of challenges arise in the predictive analytics execution process, depicted in Figure 1. These range from foundation-level deficiencies in platform and the organization to practical issues in the actual implementation of a model. For one thing, there's the not so simple matter of deploying the predictive analytics model in the real world. The scale and scope of data analytics in a production environment may require further tuning of the model as well as changes to the compute configuration. All of this takes people, who are increasingly hard to find.

IT Central Station members highlighted the following issues that can form obstacles to success with data science and machine learning projects:

- **Being unable to hire and retain data scientists**—this is one of the most serious and pervasive challenges facing organizations interested in doing predictive analytics.
- **Non-intuitive User Interfaces (UIs)**—which slow down data science project implementation.
- **Overly long project implementation times**—the pace of the predictive analytics lifecycle drags on without the right tooling, people and processes in place. Getting from the starting line to a working prototype often takes too long, and iterations are overly time-consuming. Then, getting from prototype to production may suffer from delays due to technology and process. Legal, security, governance issues and human resources problems tend to exacerbate the situation.
- **The need to involve people with diverse backgrounds**—people who don't usually know how to write code or create data models.
- **Scale and complexity**—getting bogged down in a complex predictive analytics model with an overly ambitious scope; Lacking preparation to scale and meet service level requirements.
- **Integration**—with multiple data sources, e.g. getting blocked from accessing, aggregating and exploiting data and software assets in a multi-cloud environment.
- **Vendor deficits**—a lack of expertise in setup and ongoing support for data science workloads.

Solving the Problem: 10 Tips for Visual Data Science

IT Central Station members have shared tips that help organizations overcome the challenges in effective data preparation, model development and training. With a visual data science approach based on their use of the IBM SPSS Modeler, they recommend taking advantage of tools and techniques that speed up the data science lifecycle. Many of their tips deal with empowering non-data scientists to accomplish sophisticated analytic tasks through solutions like IBM SPSS Modeler, which are designed for the business or IT generalist.

1. DEPLOY QUICKLY BY USING GUI-BASED MACHINE LEARNING ALGORITHMS

Some machine learning tools help speed up deployment of algorithms by enabling their creation through Graphical User Interfaces (GUIs) rather than a standard coding process. This is a feature of IBM SSPS Modeler admired by an IT Specialist at a small government agency. He said, "It gives you a GUI interface, which is a lot more user-friendly and easier to use compared to writing R scripts or Python, like some Anaconda type code. It makes it more open and accessible to users that are not as familiar with programming."

An Enterprise Analytics Manager at a healthcare company with over 1,000 employees chose IBM SSPS Modeler for machine learning because of its drag-and-drop algorithm building capabilities. He commented, "Most of our business analysts are non-technical, so this was attractive to them." A Founding Partner at a tech services company praised SSPS Modeler for its automated data preparation capabilities. Previously, he had many analytics jobs "stuck in Excel due to huge numbers of rows." Now, he can tackle them rapidly, noting, "The automated modeling process helps us to get going so quickly."

Analytics with visual modeling capabilities are what drove the interest of a VP, Data and Analytics at a financial services firm with over 1,000 employees. For a Senior Operations Manager at a manufacturing company with more than 10,000 employees, the best feature of SSPS Modeler was "quickness and ease of use with the guarantee of robust modeling techniques and trustworthy accuracy."

A Director of Engineering at a logistics company used SSPS Modeler to create analytical models for use cases ranging from pricing to just-in-time inventory management. He was pleased that SSPS Modeler allowed his team to put 10 models into production, quickly transforming and moving existing models into the SPSS environment. As he noted, "We saw increases in accuracy resulting from this. Therefore, we are running faster and more accurately."

> **"The ability to customize some of my streams with R and Python has been very useful to me."**

2. TAKE ADVANTAGE OF OPEN SOURCE-BASED INNOVATION INCLUDING R OR PYTHON

Open source components often accelerate implementation of machine learning models, as an Analytics Product & Services Manager at a manufacturing company with over 1,000 employees explained. He said that SSPS Modeler's "performance has been great." He then added, "I've used it for about eight years or so. [It offers] lots of flexibility. It continues to be a very flexible platform, so that it handles R and Python and other types of technology. It seems to be growing with [the] additional open-source movement out there on different platforms."

He advised, "If you're considering that open source-solution, definitely consider [SSPS] Modeler as well. Put together some kind of proposal that allows you to figure out how much time it's going to take individual people to create those models, versus being able to have an out-of-the-box solution that gets your team going more immediately."

The Quantitative Researcher at the financial services firm found SSPS Modeler "extremely easy to use" because "it offers a generous selection of proprietary machine learning algorithms with advanced tuning capabilities and integration with Python." Similarly, a Business Intelligence Manager at a manufacturing company with over 1,000 employees added further color by commenting, "I think the ease of use in

the user interface is the best part of it. The ability to customize some of my streams with R and Python has been very useful to me. I've automated a few things with that."

3. SEEK ROI BY SPEEDING UP THE END-TO-END DATA SCIENCE LIFECYCLE

A machine learning model isn't earning any Return on Investment (ROI) until it's in production and working properly. This issue figured into the comments made by a Unit Manager at an insurance company with over 1,000 employees. He described his group's capabilities with SSPS Modeler as high, adding, "They no longer waste time on modeling and algorithms, meaning they are not coding anymore. For example, segmentation projects now take one to three months, rather than six months to a year, as [they did] before."

For the Enterprise Analytics Manager at the healthcare company, the advantage of SSPS Modeler was that "it minimizes coding." This meant, "Our go-live process has been slightly enhanced compared to the previous programmatic process. There is now a faster time to production from the business end." The VP, Data and Analytics at the financial services firm also experienced a speeding up of his go-live process. He noted, "It's not just the time to go-live but it's also the process itself. The improvement in terms of performance and maintenance is also important. I would say it has saved us a lot of time, about 20% or 30% of our time."

4. EMPOWER PEOPLE WITH VARYING LEVELS OF SKILL WITH AN INTUITIVE USER INTERFACE

Predictive analytics and machine learning projects deliver business results a lot faster when they're produced by people with varying skill levels. This is a reality given how hard it is to find and retain data science professional along with experienced coders. Also, given how predictive analytics usually brings together stakeholders from multiple backgrounds, a non-technical person familiar with the business issues may actually be a better candidate to execute the project than someone who has mostly technical skills.

From this perspective, it makes sense that the Quantitative Researcher at the financial services firm would describe SSPS Modeler as "a great tool even for an individual with no or basic predictive modeling experience." A Product Team at a healthcare company said he would recommend SPSS to someone who has just started trying to run a lot of modeling. "It's a good starting point," he said. "It is very easy to use and will do the basics."

An Associate Product Manager at a financial services firm with over 1,000 employees simply said, "IBM was chosen because of usability. It's point and click, whereas the other out-of-the box-solution, or open-

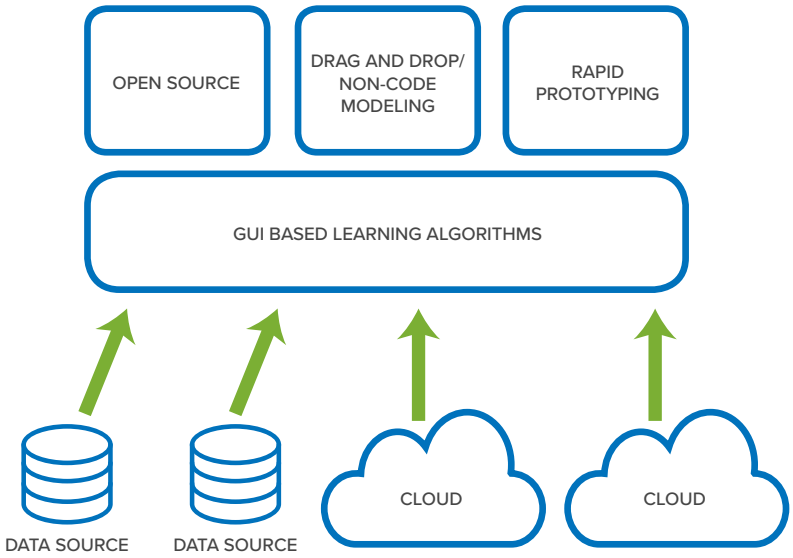

Figure 2 - Recommended solution parameters for a fast ROI with predictive analytics.

source solutions, require full-on programming and a much higher skill level." He framed the idea by reversing the perspective. He said, "If you're hiring a data scientist, you don't need IBM SPSS Modeler. If you only have an MBA who needs to be running proofs of concept, then buy IBM SPSS Modeler."

5. EXPLOIT A MULTI-CLOUD APPROACH

Machine learning models need to be able to consume data from virtually any source. Today, that means multiple cloud environments in addition to traditional on-premises databases. Figure 2 references this capability. IT Central Station members recommend selecting a machine learning solution with multi-cloud capabilities. As an Analyst at a transportation company with more than 10,000 employees described, "We have a private cloud, which is our corporate cloud. Everything is done off of a shared server." The Director of Engineering at the logistics company shared that his organization was "using a public Azure cloud. We are not deploying apps, but we are doing the analytics. We are pulling the data in with it. Then, we are writing the tables."

> **"I use it for quick prototyping. It is just a lot faster. So you do not have to write a bunch of code..."**

6. PROTOTYPE AND ITERATE QUICKLY

Rapid prototyping and iterating of machine learning models contributes to faster time to value. Users of SSPS Modeler appreciate this aspect of the solution. The Director of Engineering at the logistics company, for example, praised its "ability to quickly prototype," while the Associate Product Manager at the financial services firm liked its "rapid prototyping, [and] pre-production of models before roll out." A Clinical Assistant Professor added, "I use it for quick prototyping. It is just a lot faster. So you do not have to write a bunch of code, you can throw that stuff on there pretty quickly and do prototyping quickly."

7. INTEGRATE INTO ENVIRONMENTS TO DEPLOY REAL-TIME AND NEAR-REAL-TIME

Machine learning and predictive analytics solutions do not run in isolation. When they can be integrated into broader IT and data management environments, they can get to business value faster through real-time or near real-time deployment. The Director of Engineering at the logistics company spoke to the benefits of integration, observing that he had saved time in deployment "based off the ability to build codes quicker."

He described how, with SSPS Modeler, his team "put them into production because we have collaboration employment services, which is another analytic solution from IBM, so we are able to productionalize the models and manage the models from this environment." He concluded, "Altogether, this saves us a lot of time versus if we want a programmatic solution and had to have developers write C# and Java around it. Overall, it is a huge increase to time savings."

Other IT Central Station members benefiting from SSPS Modeler's integration capabilities include the IT Specialist at the small government agency, who said, "We have integration where you can write third-party apps. This sort of feature opens it up to being able to do anything you want." The Director of Engineering at the logistics company praised SSPS Modeler for its "integration into all the existing environments."

Integration with other business intelligence tools is of particular importance, as the Analyst at the transportation company shared. "We are putting seven machine learning models in production to start. We may expand up to 10," He said. "This is real-time, as we are pulling data out of Cognos BI server every morning. We manipulate and reload the data throughout the day based on parameters that come in from the field. Then, that gets put back into the system and refreshed for the next day."

The IT Specialist at the small government agency also discussed the value of BI integration in terms of his future machine learning plans. He said, "We're doing real-time right now, but we are doing batch once we get the server product up and going. In terms of models, we are getting it off the ground. We have been using it for about six months, and we have been just playing with getting our models up and going, so we actually have the whole pure data and Hortonworks analytics products that we are going to be deploying in the analytics environment. That's

where our server product will go. Then we will have all of the governance pieces in place to start doing production deployment. So, we are almost there."

8. START SMALL AND SCALE THE SOLUTION UP AND OUT

IT Central Station members advise new adopters of machine learning to start small. Even with intuitive, GUI-based tools, the analytical processes involved is sufficiently challenging to make overly ambitious early projects an unwise idea. As an Analyst at a transportation company with over 1,000 employees put it, "Give it a try. Start with a proof of concept and see where it leads. Right now, I think we have about five or six different machine learning proofs of concept, using real-time data. We're running them on Bluemix / IBM Cloud." The Founding Partner at the tech services company advised, "Do not dive into the server directly. It is very hefty for just doing calculations that can already be done by SQL Server R or Oracle. Maximize the utilization of the desktop tool first."

> **"Start with a proof of concept and see where it leads. Right now, I think we have about five or six different machine learning proofs of concept, using real-time data."**

9. LEVERAGE ONLINE DOCUMENTATION

Machine learning practitioners are becoming members of a large community. Many are learning that others have previously tackled the same kinds of difficult predictive analytics challenges they are facing now. With the right vendor, useful solution ideas show up in documentation. As the Quantitative Researcher at the financial services firm described, "[With] the very detailed online documentation and examples that IBM SPSS Modeler provides, even a novice employee can start using the tool and become productive in a short period of time."

10. LOOK FOR PROVEN EXPERIENCE AND EXPERTISE IN A VENDOR

Vendor choice, important for success in any IT scenario, is distinctly relevant for the successful adoption of machine learning. With the paucity of experienced data scientists and the subjective, complex nature of the terrain, the right vendor can make a major difference in predictive analytics outcomes. To this point, the Director of Engineering at the logistics company shared, "I chose IBM SPSS because of their experience with the solution, what they brought to bear, and their relationships." The Analyst at the transportation company added, "The most important criteria when selecting a vendor [is] ease of use. They should be able to handle our unique situation. We have many branches with many moving parts, and also a lot of internal customers."

Further to the theme of vendor experience, the Business Intelligence Manager at the manufacturing company explained, "What's most important when selecting a vendor is the proven practice of the product. [It's useful] knowing that the product has had success for numerous other customers in the past for similar use cases, for similar types of customers. I think knowing that there are a variety of partners out there with expertise in the product is a very strong selling point for me. I don't like going to things where I can't get help if I get stuck."

CONCLUSION

Getting to success with machine learning and predictive analytics requires a mix of people, processes and tools. IT Central Station members shared their experiences with the IBM SSPS Modeler to highlight tips for getting the most out of an investment in machine learning. Their insights emphasize the benefits of solutions that enable non-coders and non-data scientists to build and deploy models for data science projects for enterprise deployment. More broadly, they recommend solutions that make it possible for machine learning projects to advance quickly by streamlining the processes of data preparation and model creation.

According to IT Central Station members, effective machine learning and predictive analytics flow from visual data science solutions that deploy quickly through the use of GUI-based machine learning algorithms. The goal is to prototype and iterate quickly. Open source compatibility, especially with R and Python, further accelerate the modeling processing. Integration with a variety of environments, coupled with a multi-cloud approach, facilitates access to data resources in multiple locations. With an experienced vendor and the right solution, it is possible to derive desired business outcome and realize strong ROI with data science and machine learning in a business context.

ABOUT IT CENTRAL STATION

User reviews, candid discussions, and more for enterprise technology professionals.

The Internet has completely changed the way we make buying decisions. We now use ratings and review sites to see what other real users think before we buy electronics, book a hotel, visit a doctor or choose a restaurant. But in the world of enterprise technology, most of the information online and in your inbox comes from vendors but what you really want is objective information from other users. IT Central Station provides technology professionals with a community platform to share information about enterprise solutions.

IT Central Station is committed to offering user-contributed information that is valuable, objective and relevant. We validate all reviewers with a triple authentication process, and protect your privacy by providing an environment where you can post anonymously and freely express your views. As a result, the community becomes a valuable resource, ensuring you get access to the right information and connect to the right people, whenever you need it.

www.itcentralstation.com

IT Central Station does not endorse or recommend any products or services. The views and opinions of reviewers quoted in this document, IT Central Station websites, and IT Central Station materials do not reflect the opinions of IT Central Station.

ABOUT IBM SPSS MODELER

IBM SPSS Modeler is a leading visual data science and machine learning solution. It helps enterprises accelerate time to value and desired outcome by speeding up operational tasks for data scientists. Leading organizations worldwide rely on IBM for data discovery, predictive analytics, model management and deployment, and machine learning to monetize data assets. IBM SPSS Modeler empowers organizations to tap data assets and modern applications with over 40+ out of the box algorithms and models, suited for hybrid, multi - cloud environments with robust governance and security posture.

IBM SPSS Modeler empower organizations to:

- Take advantage of open source based innovation including R or Python
- Empower data scientists of all skills programmatic and visual
- Exploit hybrid cloud approach – on-prem, public or private clouds
- Start small and scale to enterprise

IBM SPSS Modeler is available by subscription, perpetual license or as part of IBM Data Science Experience. To learn more, please visit: https://www.ibm.com/products/spss-modeler