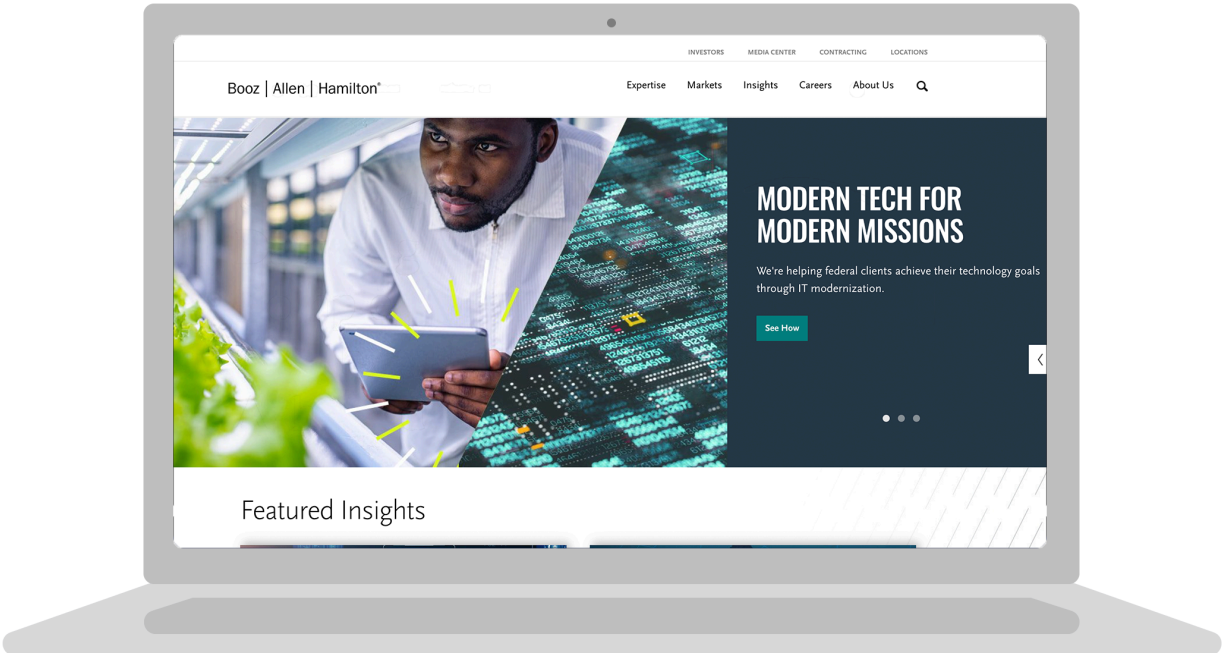
Case Study: Machine learning model unearths resourcing insights for Booz Allen Hamilton

Company background

For more than 100 years, government, military, and business leaders have turned to Booz Allen Hamilton to solve their most complex problems. Their experts in analytics, digital solutions, engineering, and cyber work together to find solutions that help organizations transform. They are a key partner on some of the most innovative programs for governments worldwide and are trusted by these governments' most sensitive agencies. They work in partnership with clients, using a mission-first approach to choose the right strategy and technology to help them realize their vision. With global headquarters in McLean, Virginia, Booz Allen Hamilton employs nearly 29,300 people globally as of June 30, 2022, and had revenue of $8.4 billion for the 12 months ending March 31, 2022. To learn more, visit www.boozallen.com.



The challenge

Booz Allen Hamilton's work comes from contracts awarded by federal government and commercial clients. To get this work, the company bids on contracts. The bidding process involves reviewing the job requirements and submitting a proposal for a solution and a cost to implement it. The organization that solicited the bids will then review these proposals and award a contract to the company whose bid best suits their needs and budget. Upon award, Booz Allen must meet those requirements and deliver the best possible solution to their client. For a company of its size, the uncertainty that comes with bidding and winning work can make staffing and resource planning difficult. With this machine learning model they are able to better understand which contracts they'd be more likely to win and therefore understand the workforce and skills needed to meet its future demand.

Step 1: The stakeholders pave the way

Even though the model would only be used internally at Booz Allen Hamilton, it still involved numerous stakeholders, all of whom needed to work together to develop and deploy the model. At large organizations, this can take days, weeks, or even months. For this model, the stakeholders include:

- **Corporate Finance team:** This is the team that will use the model. They have domain expertise of the model's applications and use cases. They dictate what the model should do, and they know what data is relevant and available to build it.
- **Enterprise Data Science team:** This is the team responsible for training, validating, and deploying the model. Throughout model development, they work closely with the model end-users (the Corporate Finance team) to collect feedback, explore use cases, and validate results.

- **Enterprise Platforms and Engineering team:** This is the team responsible for storing and maintaining the data in an enterprise data lake (a large repository of structured and unstructured data). Large organizations typically have so much data that there is a team tasked with managing all of it. The Platforms and Engineering team also is responsible for setting up the necessary computing platform and infrastructure where the model is developed and deployed.

Step 2: Develop the model

To provide more insight to leadership and decision-makers, the Data Science team created a tree-based classification model that predicts the likelihood of winning a contract.

Target variable and model output: The target variable was binary: whether or not a submitted bid won Booz Allen Hamilton the contract. The model's final output is the probability of winning a given contract.

Algorithms considered: The Data Science team explored a number of different algorithms for the model, including logistic regression, support vector machines, decision tree, and random forest. A random forest model was ultimately selected as the champion solution.

Class balance: There was a minor imbalance between the classes of the target variable. Both upsampling and downsampling were tried, but ultimately neither provided any substantial lift over the baseline.

Evaluation metrics: The model was evaluated based on its performance with respect to four different metrics:

1. Area under the ROC curve (ROC/AUC)
2. F1 score
3. Accuracy
4. Log-loss

Splitting the data: The data comprises several full years of historical data as well as resolved bids from the current fiscal year. The data was split into training and test sets.

- **Training data:** Several years of historical data and 50% of resolved bids from the current fiscal year
- **Test data:** The other 50% of resolved bids from the current fiscal year

Model training & tuning: The Data Science team tuned four main hyperparameters using 5-fold cross-validation:

- **Max samples:** The number of observations sampled with replacement
- **Max features:** The number of features to consider when looking for the best split
- **Number of trees:** The number of base learners grown in the ensemble
- **Tree depth:** The level to which each tree is allowed to grow

Feature selection: Of the 250+ initial features, approximately 40 features of varying types (numerical, categorical, and Boolean) were selected for use in the final model based on their relative importance.

Model selection: The final model was a random forest model that was selected based on its performance on the test data as indicated by the four metrics listed above.

Step 3: Deploy the model

Once the Data Science team finished building the model, they drafted a report to present to the Corporate Finance team. This report contains details of the model development and validation process, including model architecture, performance results, important features, and relevant visualizations that support their conclusions. One of the most important criteria for model approval and deployment is that the model provides an improvement over existing methods of estimation. In this case, the final model proved to be 12% more accurate than existing methods.

After receiving buy-in from the Corporate Finance team, the Data Science team, with support from the Platforms and Engineering team, worked to deploy the model so end users can begin using its predictions. Enterprise-level data science is not performed on a personal laptop. The volume of data, computing requirements, and risk are all too great for this. The model must be developed and deployed on a platform that is powerful, reliable, and secure enough to support it. The Data Science team developed and deployed this model leveraging the enterprise data lake infrastructure, which includes a notebook-based model development environment as well as a platform to streamline machine learning development, management, and deployment.

Key takeaways

Below is a summary of the main insights from this case study.

- Machine learning solutions are transforming the way businesses operate, and these solutions are not all customer-facing; they're also used to aid in internal processes and decision-making.
- Development of machine learning solutions at large businesses involves many stakeholders with diverse responsibilities and fields of expertise.
- Many of the algorithms taught in this certification—including logistic regression, decision trees, and random forests—are used by the biggest companies in the world, and their modeling process closely resembles that which is presented in this course's notebooks.
- Even when a model performs well, its use must still be carefully explained and justified before deployment. It is not enough to simply build a model that scores well.