Best Practices
for Planning Your Next Machine Learning Project

Arthur

Best Practices for Planning Your Next Machine Learning Project

According to Gartner, only 53% of machine learning models convert from prototype to production. What can you do to ensure that your next machine learning model is not part of the 47% of machine learning projects that fail to go into production?

Success of your next AI project depends on understanding more than the technology, models and data. An O'Reilly survey discovered the biggest obstacles to enterprise ML adoption are institutional support and difficulties in identifying relevant business use cases.

Here are 8 essential steps for planning your team's next project:

What is Arthur?

- Arthur is the AI performance company. We work with enterprise teams to **monitor, measure, and improve machine learning models** for better results across accuracy, explainability, and fairness.
- Arthur's **research-led approach to product development** drives exclusive capabilities in computer vision, NLP, bias mitigation, and other critical areas.
- Arthur has partnered with Fortune 50 enterprises across financial services, healthcare, retail, and tech to **help scale their ML monitoring and validation capabilities and improve AI performance.**

Here are some best practices that leading enterprises follow to advance their ML operations to drive business impact.

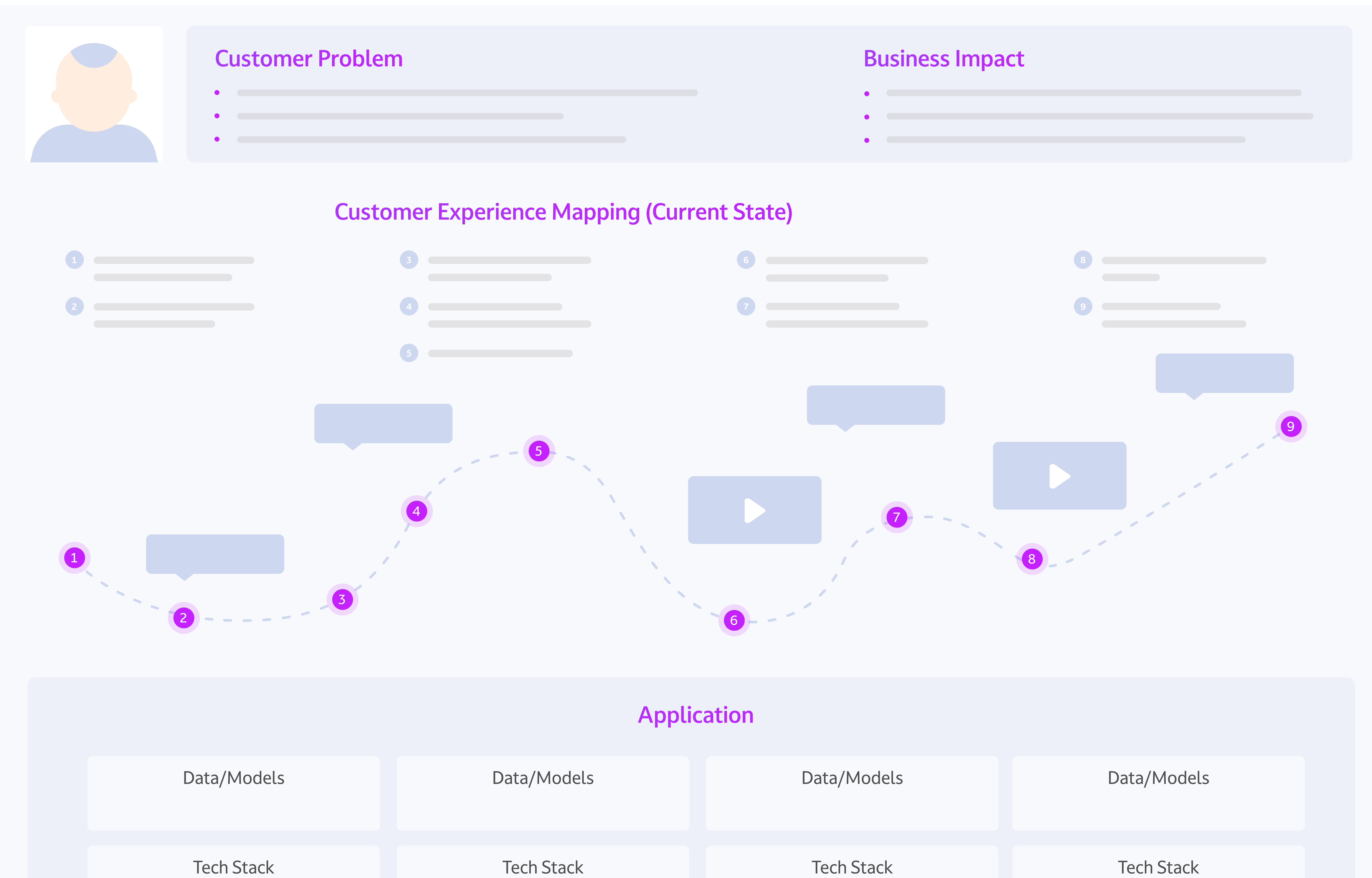
STEP 1

Discover How Data Science Drives Business Value

Identify the customer-centric problem and business value of an ML model

Many enterprises don’t have a clear roadmap or vision for how AI can solve business challenges. Business impact metrics are downstream KPIs that are impacted by ML models. For example, how would an ML model impact customer trust, satisfaction, churn, and revenue?

Below is an easy-to-use framework for architecting how your ML model influences product, operations, customer experience and business teams.



Start small and simple

Identify one customer-centric problem and how it directly impacts the business (churn, NPS/CSAT, CLV, etc.).

Map customer journey

From initial touchpoint to the last touchpoint, detail the customer experience.

Diagram the back-end

Inventory datasets, model & tech stack dependencies that inform specific moments in the customer journey.

STEP 2

Evaluate the Enterprise Value of a Single ML Model

Assess ML models within a wider enterprise context

Congratulations! You've identified a customer problem, the business impact of it, mapped the customer experience and application/data/models powering the predictions for the application.

Now it's time to assess whether it's worth your team's time, resources and money to scope and undertake a project focusing on a single ML model. **If the answers are a resounding yes, proceed to Step 3.** If the answers are predominately no, your team may opt to go back to the drawing board and select another customer problem requiring a new journey remapping.

Customer Problem	Is the customer problem material: **widespread or edge case?**	Is this customer problem a **quarterly priority** for operational leaders and C-Suite?
Business Impact	Is the business impact of this model **understandable and clearly defined?** Is there sufficient dollar or % impact?	Does the business have **resources to participate** in the project based on the timeline?
Data/Machine Learning	Do you have the **required datasets** that you need to train the model & measure what you want?	Do you have access to the datasets? **What's the quality of the data?**
Technical Implementation	Do you have the **cross-functional resources** to deploy and test the model?	Can you **execute the model in a timely manner** to meet business/operation needs?

STEP 3

Seek Stakeholder Involvement & Buy-In

Engage with stakeholders who have an interest in ML model performance success across the company

Use this checklist or create your own to identify and nominate cross-functional stakeholders inside your company, and/or external consultants who are chiefly responsible for the customer journey you just mapped across omnichannel experiences; think product, operations, analytics, ML Ops and data science.

Scoping questions to ask

Customer Experience & Business Metrics

→ What are the business metrics for this customer journey segment? **Who owns those business metrics?**

Product Owners

→ What product/application features does the ML model **directly impact?**

Enterprise Database Architecture & Infrastructure

→ What is the **database architecture and infrastructure** underlying the models?

→ **What APIs are necessary** so stakeholders responsible for this journey segment can receive relevant alerts and notifications when there's a post-production issue to reduce risk?

Project Roles

→ Who are the operational, product and ML/Data Science **leaders who are responsible and accountable** for this customer journey segment?

Data Science/ML Models

→ What is the **# (quantity) of models and inferences** that informs this customer journey segment?

→ Is there a centralized view, source of truth for all the models that impact this customer journey segment?

→ What are the **inputs and outputs** for this ML model?

→ What are the **data science metrics** used to evaluate model performance (industry standard vs. custom or hybrid?)

→ **How often** are the datasets informing this model **reviewed, scrubbed or updated?**

STEP 4

Map Model Variables to KPI Metrics

Linking Inputs and Outputs to Business

To assist your in-house analytics team in understanding technical and business metrics you'd like tracked within a business intelligence dashboard (Tableau, etc.) to evaluate your project's progress and success, take some time to create a ML Ops to Business relationship chain. We've included a few examples below using some industry inputs:

	Input	Output	Technical Metric	Business Metric
Financial Services	Transactions	Fraud Transactions	F1, Precision, Recall	Revenue Loss, Fraud To Sales Ratio, Etc.
Retail/E-commerce	Purchase History	Recommendations	Mean Average Precision At K	Uplift In Average Revenue Per User, Or Number Of Items Added To Cart, Etc.
Healthcare	Clinical Symptoms	Diagnosis	F1, Precision, Recall	Total Cost Of Care Reduction/ Savings, Improved Health Outcome, Uplift In Medication Adherence.

Linking inputs to outputs and then tying technical metrics to business metrics is a critical step. Having the ability to query and export any metric from the Arthur platform into your own BI platform is essential at this stage for understanding how your model impacts business value and influences KPIs.

You can use Arthur's robust API service to query and integrate all metrics (model performance, data drift, bias, explanation) into applications for task automation, product innovation and ROI evaluation.

Scope Your ML Project & Gain Approval

Collect inputs to inform ML project scope/framework

Next, it's time to inventory and scope the project's inputs and outputs, as well as set ML Ops goals and metric benchmarks. Lean on a program or project manager to help you roadmap deadlines, risks and dependencies.

The likelihood your project will be approved by procurement and/or line of business stakeholders increases when you can show you have a performance measurement plan in place with Arthur to evaluate the effectiveness of your next ML project.

Problem Definition

- → What output do you desire to predict? What are key factors for being able to predict outputs?
- → What input data do you have to inform the algorithm?
- → What and where are the input datasets located?

Performance Measurement

- → Do you have existing benchmarks for comparison?
- → How will you measure prediction accuracy?
- → What is the minimum % for ground truth?

Timeline

- → When do you need to start the project?
- → What is the deadline for launching the model into production?
- → What is the timeline to start proving success in production?

STEP 6

Prove Your Model Works

Perform 2 week POV evaluation & validation

Keep these principles in mind when performing model testing.

Accuracy

→ Is the model effective for the desired task at hand? Quantify using the appropriate model evaluation metrics (e.g. AUC, RMSE, F1, etc).

Robustness

→ Is the model stable (e.g., explicitly test for drift, noise, bias)?

Interpretability

→ Does the model predict outputs as it should? How do input variables contribute to the output?

Reproducibility

→ How can you make your results reproducible when your model changes due to parameter adjustments, retraining, or new data?

→ Make sure all model code is versioned and reproducible; all training and testing data is stored; all validation experiments and results are stored (along with hyperparameters).

Testing a full model requires a significant commitment, both in team resources and time. Instead, test micro model components or utilize a smaller dataset.

Split your data into a test set (~15% of total data set) and a validation set (~ 15% of total data set).

To test the accuracy of a model, utilize the test set. To avoid possible data bias, make your test sets available to the model building team (aka the maker team).

To further reduce bias in the model, make your validation set only available to a different team that validates the model metrics against the validation set (aka the validation team).

When working with test and validation data sets, experimentation is essential to ML model optimization.

Arthur's model versioning empowers you to optimize your ML models by easily comparing models across versions (new model iterations, A/B test, shadow-mode). Furthermore, you can compare two versions of your model by schema differences or metric differences (performance, inferences accuracy, bias, data drift). This empowers your team to easily manage model changes over time and responsibly test model versions before moving to operations. Best of all, post-production, you can monitor all your models in one place with our highly scalable microservices architecture. With Arthur, you can ingest millions of data points and deliver insights quickly.

STEP 7

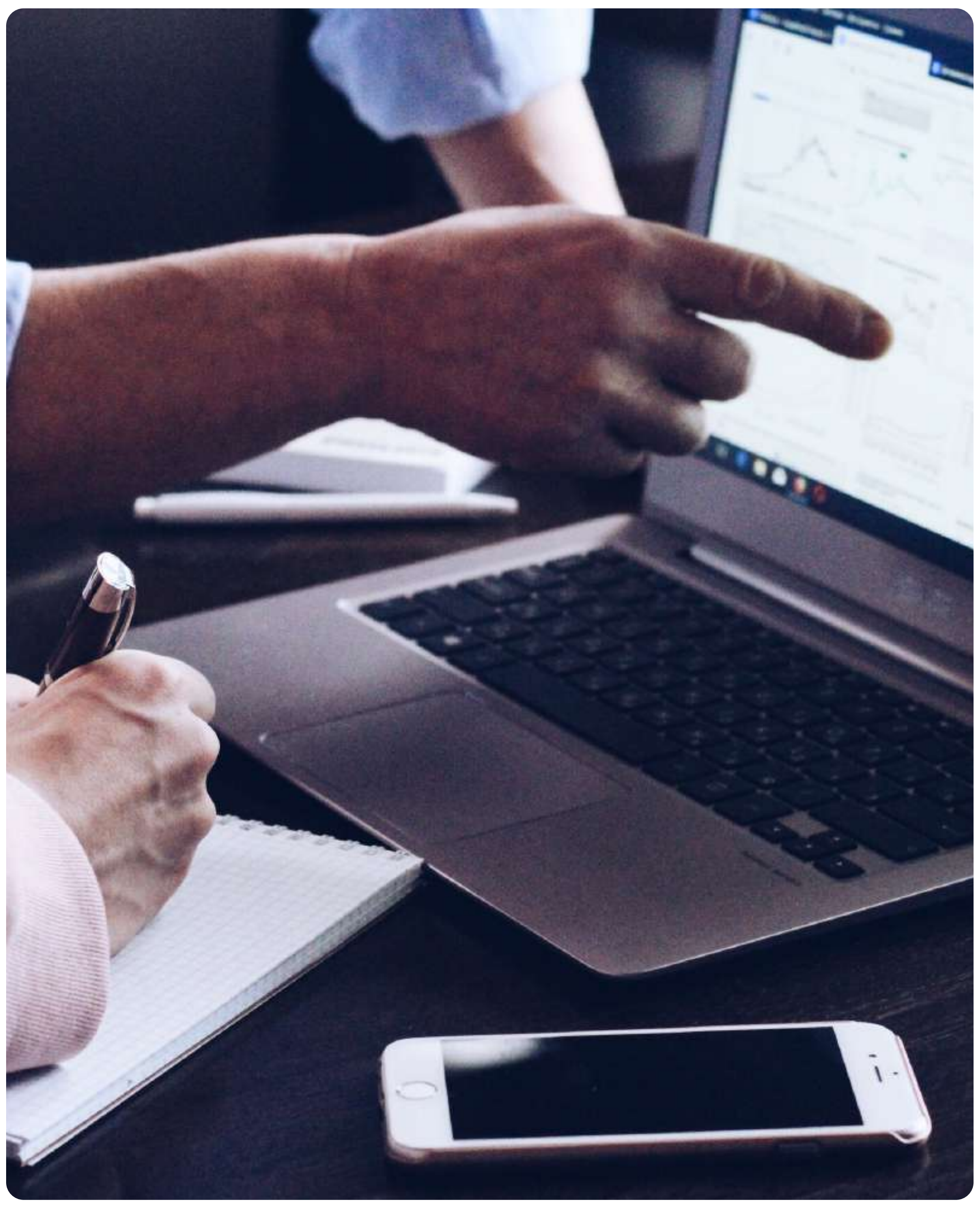
Build a Culture of AI Transformation & Innovation

Share out ML project results with xfn team

Clear communication and collaboration between data science/MLOps and business teams is essential to create a center of AI excellence and culture of data-driven innovation.

Celebrate ML project success internally via intranet posts, quarterly meetings, webinars, lunch and learns and more. If your ML model project is not proprietary, get PR/marketing permission to discuss ML model success publicly in podcasts, meetups, workshops or upcoming conferences.

By using Arthur to measure ML Performance and continually monitoring models post-production, you can better help convey the cross functional business value add and ROI of your project.



STEP 8

Establish Repeatable Workflows for ML/AI Model Governance

Operationalize new or adjusted models to continually measure & improve performance

Release to Production	Monitor & Iterate	Process	Quarterly Reporting	Govern & Audit
Models evaluated post-production.	Continually monitor models to trigger revisiting data sets, retraining or retiring ML model actions.	Set repeatable processes and calendar cadences for revisiting models when KPIs change and/or customer experience journey shift.	Share out post-model optimization impact on business performance and customer experience.	Partner with risk/ compliance to understand audit log and versioning documentation to satisfy pre-audit readiness, independent audit or post-audit remediation.

For performance and governance, you need a central destination to monitor models post-production. Arthur is the only AI performance solution that supports model types including multi-class, multi-label, and regressions, allowing you to measure and optimize all model types in one place.

In addition to Tabular, Arthur measures and improves NLP and CV models across both monitoring and validation workflows.

For CV models, you can leverage the platform to perform object detection, find anomalies in incoming images plus also monitor for drift and bias, and provide local explainability. For NLP models, find anomalies in incoming text, monitor for drift and bias, and provide local explainability.

To ensure your model is not resulting in unintended bias, you can get actionable insights into how your model treats different population groups and use Arthur's proprietary bias mitigation techniques to reduce business risk and prevent discrimination.

Arthur empowers teams to set fairness thresholds to uphold your organization’s data code of ethics and algorithmic risk committee charter. Get notified via alerts instantly if there are any problems to proactively manage model risk.

Tips

- → Start small with one use case, get a model win and then scale ML model projects across additional customer journey segments.
- → Look for high-impact, low-effort use cases across the AI customer journey for ML model experimentation.
- → Remember, ML will not solve all problems. There are fundamental differences between ML model performance and business performance metrics of success. Good model performance does not always guarantee successful business performance.