

Dataiku Machine Learning Sample for Car Images Classification to produce automobiles

Background

Chilly, a small city with a limited number of cars (80 in total), is looking to optimize automobile production to meet future demand. As a new car manufacturer, we are starting our business from this small city. Our goal is to automate the classification of vehicles entering the city and determine which type—Sedan or SUV—is more prevalent. This data-driven approach will allow us to manufacture the vehicle type that has higher demand, ensuring profitability and efficiency.

Problem Statement

Currently, there is no automated system to classify vehicles in Chilly. A manual process would be inefficient, time-consuming, and prone to human errors. Without data-driven insights, the company risks producing vehicles that may not align with market demand as the automotive industry relies heavily on efficient inventory management, customer segmentation, and pricing optimization. The capability to automatically classify cars into sedans and SUVs presents significant opportunities across various business applications. Dealerships and car manufacturers can analyze market demand and optimize inventory to align with customer preferences. Understanding which vehicle type is more popular in specific regions enables data-driven decisions, and pricing strategies can be adjusted based on real-time classification data reflecting supply and demand trends.

Business Case and Market Justification

The global automobile market is increasingly data-driven, with companies like Tesla, Ford, and Toyota using AI for production planning. However, new manufacturers in small markets lack real-time demand insights.

For Chilly's emerging auto market, identifying the dominant vehicle type is critical for:

1. **Production efficiency** - Avoid manufacturing models that don't sell.
2. **Cost reduction** - Minimize resource waste on unpopular vehicles.
3. **Strategic sales targeting** - Market based on consumer preference.

Market Impact of AI-driven Classification

1. **SUVs** dominate sales in regions with rugged terrains and extreme weather.
2. **Sedans** remain popular in urban centers due to fuel efficiency.
3. AI-driven classification allows real-time trend analysis, ensuring scalable, data-backed manufacturing decisions.

Proposed Solution

We will develop an AI-powered image classification model that automatically detects and categorizes cars as either **sedan** or **SUV**. This system will be trained using images captured from within Chilly, allowing us to determine the predominant vehicle type. The model will then be deployed at key entry points to classify any new vehicles entering the city in real time.

Data Collection

To build the classification model, a dataset of the total cars in the city (80 images) were collected, labeled, and preprocessed. These images were sourced from real-world street photography, with manual labeling performed to ensure ground truth validation.

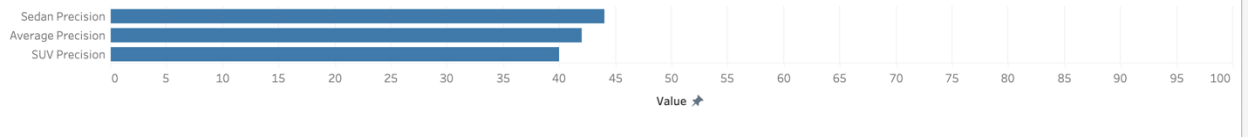
Data Processing

Data preprocessing included image augmentation techniques such as rotation and flipping to improve model generalization. Bounding boxes were drawn to detect and classify vehicles, ensuring accurate feature extraction. The dataset was split into an 80 percent training set and a 20 percent test set to facilitate model evaluation.

Model Training & Performance Evaluation

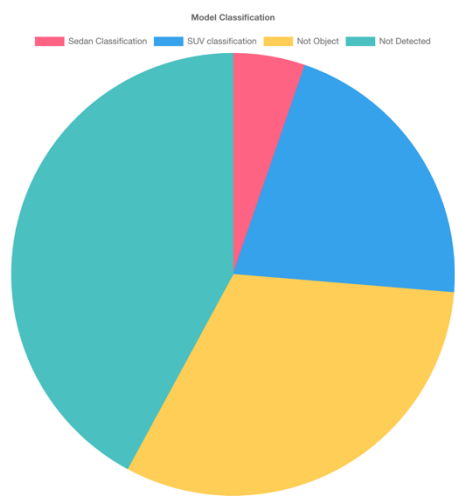
The classification model was developed using object detection techniques within Dataiku, employing a convolutional neural network (CNN) to recognize patterns in vehicle shape, size, and structure.

Model performance was evaluated based on average precision and recall metrics.



- **Average Precision:** 41.95% at an IoU threshold of 0.5.
- **Sedan Precision:** 43.81%.
- **SUV Precision:** 40.10%.
- **Sedan Classification Accuracy:** Correctly detected 1 out of 14 instances.
- **SUV Classification Accuracy:** Correctly detected 4 out of 14 instances.
- **False Positive Rate:** 6 instances misclassified as non-objects.
- **Precision-Recall Analysis:** Precision started at 80% but decreased with higher recall values.

Confusion Matrix Insights



The confusion matrix revealed key insights into the model’s classification accuracy. Sedans were correctly detected only once out of fourteen instances, indicating significant misclassification issues. SUVs demonstrated slightly better classification accuracy, with four correct detections. A notable challenge was the high rate of false positives, where six instances were misclassified as non-objects. The precision-recall curve further highlighted the model’s limitations, with precision starting at 80 percent but decreasing as recall increased. This suggests that many vehicles were not being detected accurately, necessitating further refinement.

Several areas for improvement were identified to enhance model performance. Increasing the dataset size to at least 200 images as the number of cars start to get drove in this city would improve generalization and classification accuracy. Balancing the dataset by ensuring an equal number of sedan and SUV images would help address class imbalance issues. Enhancing bounding box annotations and adjusting confidence thresholds would also minimize false detections and improve overall precision.

Key Performance Indicators (KPIs) for Business Impact

1. **Model Accuracy Improvement:** Increase average precision from 41.95% to at least 70% within six months.
2. **Sedan vs. SUV Classification Ratio:** Ensure balanced classification with less than 10% variation in detection accuracy.
3. **Data Volume Growth:** Expand dataset to at least 500 labeled images within one year.
4. **Reduction in False Positives:** Lower misclassification rate from 6 instances to under 2 per batch of 50 classifications.
5. **Inventory Optimization:** Align stock levels with classification trends, reducing unsold inventory by 20% within six months.

6. **Customer Segmentation Accuracy:** Achieve at least 80% alignment between classified vehicle type and customer purchase preference based on sales data.
7. **ROI on AI Implementation:** Track cost savings from AI adoption, aiming for a reduction in production waste and unsold inventory costs.

Business Insights & Actionable Steps

1. **Gradual Model Deployment:** Due to model performance limitations and the limited number of cars in the city, the classification system should first be used for trend analysis rather than direct production decisions. This allows for iterative improvements in model accuracy before committing significant resources to manufacturing. In addition, periodic retraining of the model using updated datasets will ensure improved classification precision over time. The deployment should start with a small-scale pilot in key locations, followed by phased implementation across larger regions as accuracy improves.
2. **SUVs Are More Detectable:** Since the model identifies SUVs more effectively than sedans, initial market demand appears skewed towards SUVs. Thus, manufacturing efforts should prioritize SUV production while improving model accuracy for sedans. Additional data collection and annotation of sedan images can help balance classification accuracy. A dedicated task force should work on refining the model's ability to detect sedans using advanced feature extraction techniques. Meanwhile, market surveys should be conducted to validate whether SUV dominance is due to actual demand or model bias.
3. **Strategic Inventory Planning:** Given the high rate of undetected sedans, an adaptive inventory approach should be taken. A small batch of sedans should still be maintained in stock while SUVs are prioritized. This ensures that production aligns with real-time classification insights without completely neglecting sedan availability. Advanced inventory management systems integrated with AI classification data should be implemented to dynamically adjust production schedules. Furthermore, pre-orders and customer preference surveys should be leveraged to validate market trends before finalizing manufacturing quotas.
4. **Targeted Marketing & Sales Strategy:** If SUVs dominate classification results, marketing campaigns should highlight SUVs' advantages to reinforce consumer preference. However, incentives or discounts can be offered for sedans to balance supply-demand dynamics. Digital marketing strategies should incorporate real-time classification insights, allowing for region-specific advertising campaigns. Additionally, dealership engagement should be enhanced by providing data-driven insights into local market trends, enabling personalized customer interactions and optimized sales approaches. Test drive promotions and trade-in offers can be introduced to increase sedan visibility and appeal.
5. **Continuous Improvement & Expansion:** As the model's accuracy improves, classification insights should be integrated into a broader business strategy for expansion beyond Chilly, ensuring scalable and adaptive manufacturing. Future iterations should include the integration of external datasets from larger cities to improve model robustness. Expansion strategies should focus on creating strategic partnerships with regional distributors to facilitate seamless market entry. Additionally, exploring adjacent applications such as fleet management and urban mobility analytics can create new revenue streams. A dedicated R&D division should oversee long-term model enhancement, ensuring adaptability to evolving automotive trends and consumer preferences.

Conclusion

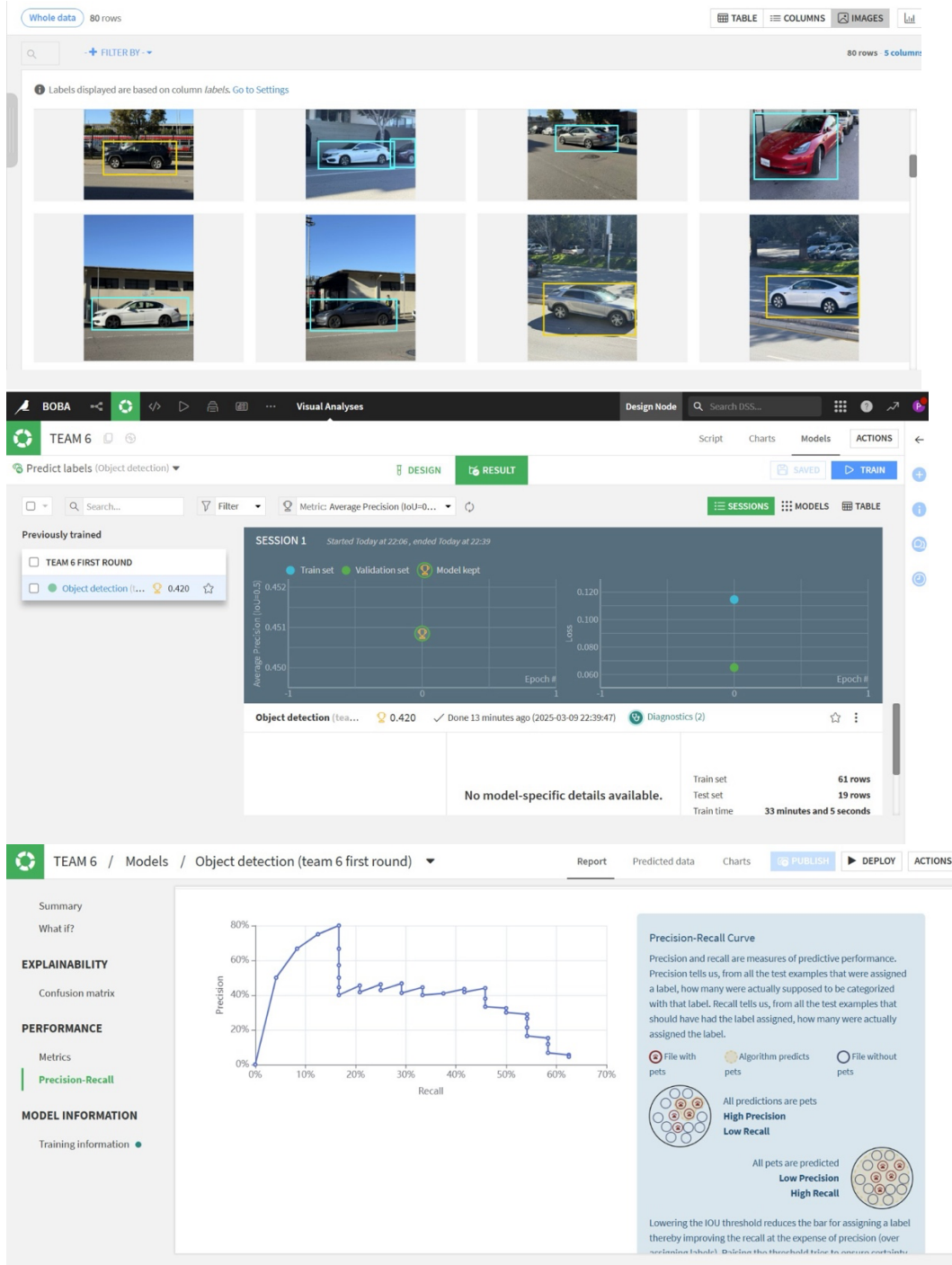
As a new car manufacturer in Chilly, leveraging AI-driven image classification allows us to make data-backed production decisions. With SUVs being more prevalent due to the city's rugged terrain and extreme weather conditions, the focus should initially be on scaling SUV manufacturing while monitoring sedan demand trends. A phased approach to production, marketing, and inventory planning will ensure optimal resource utilization and profitability. Over time, as the model improves and the dataset grows, we can refine our classification accuracy and expand our business strategy to other markets, making our approach both sustainable and scalable.

APPENDIX

Goodall, N. J. (2016). Can machine learning improve automotive safety? IEEE Intelligent Transportation Systems Magazine, 8(1), 20-25.

Thrun, S. (2010). Toward robotic cars. Communications of the ACM, 53(4), 99-106.

Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.



TEAM 6 / Models / Object detection (team 6 first round)

ReportPredicted dataChartsPUBLISHDEPLOYACTIONS

SummaryWhat if?

EXPLAINABILITY

PERFORMANCE

MODEL INFORMATION

Confusion matrix

IOU0.510.50Confidence score010.35REVERT TO OPTIMAL





Ground Truth	Predicted		
	SEDAN	SUV	Not Detected
SEDAN	1	0	13
SUV	0	4	6
Not an object	0	6	0

19 Images

Sort by Confidence | IOU

Ground Truth == Any class

Predicted == Any class



TEAM 6

ScriptChartsModelsACTIONS

Predict labels (Object detection)

DESIGNRESULT

SAVEDTRAIN

Search...Filter

SESSIONSMODELSTABLE

Name	Trained	Train time	Average Precision (IoU=0.5)	Average Precision (IoU=0.75)	Average Precision (all IoUs)	
Object detection (team 6 first round)	2025-03-09 22:06:42	33m 5s	0.420	0.048	0.092	

Performance metrics

Filter class...

	Average Precision (IoU=0.5)	Average Precision (IoU=0.75)	Average Precision (all IoUs)
All classes	0.4195	0.0476	0.0919
SEDAN	0.4381	0.0952	0.1167
SUV	0.4010	0.0000	0.0671