By Ashwini Abhang IIITB

MLOps DL- Case Study

Q1. System design: Based on the above information, describe the KPI that the business should track.

Answer 1: - Among the key performance indicators (KPIs) that the Carsdepo.com business should monitor, the following are some of the most crucial KPI's divided based on the following parameters: -

1. **Output quality/ Model Performance KPIs: -**
    - **Accuracy of damage detection:** This is the most crucial metric, as it determines the model's ability to reliably detect scratches and dents in photos. A high accuracy score will ensure that the estimated resale value of a vehicle is accurate.
    - **Precision and recall:** Precision is the ratio of the number of true positive detections to the sum of true positive and false positive detections. Recall is the ratio of the number of true positive detections to the sum of true positive and false negative detections. Precision and recall are essential metrics to monitor to ensure that the model does not overlook any damage or indicate erroneous damage.
    - **False positive rate:** This is the proportion of false positive detections to total detections. A high proportion of false positives might cause an automobile to be overpriced, which can negatively affect the consumer experience.
    - **False negative rate:** This is the proportion of false negative detections to total real damages. A high percentage of false negatives might result in the underpricing of a vehicle, resulting in a loss for the seller.
    - **Speed of detection:** The time it takes for the model to detect damage in an image is also an essential element, as it affects the user experience as a whole. A quick detection time will ensure that the consumer does not have to wait for the findings for an extended period of time.

2. **Customer-focused/ Experience KPIs: -**
    - **User satisfaction:** Customers might be asked to review their experience with the damage detection system via surveys or feedback forms. High customer satisfaction levels suggest that the model provides accurate results and enhances the consumer experience as a whole.
    - **Adoption rate:** Adoption rate is the proportion of customers who utilize the damage detection system vs those who do not. A high adoption rate suggests that users find the technology beneficial and choose to utilize it.

3. **Business-focused/ Impact KPIs: -**
    - **Cost savings:** If the damage detection system can eliminate the requirement for manual inspections, the firm can realize significant cost savings. This can be determined by comparing the cost of manual inspections against that of the damage detection system.
    - **Scalability:** It is essential to monitor the scalability of the damage detection system as the firm expands. This can be determined by monitoring the system's processing time and resource consumption as the number of photos rises.

By Ashwini Abhang IIITB

Q2. System design: Your company has decided to build an MLOps system. What advantages would you get from building an MLOps system rather than a simple model?

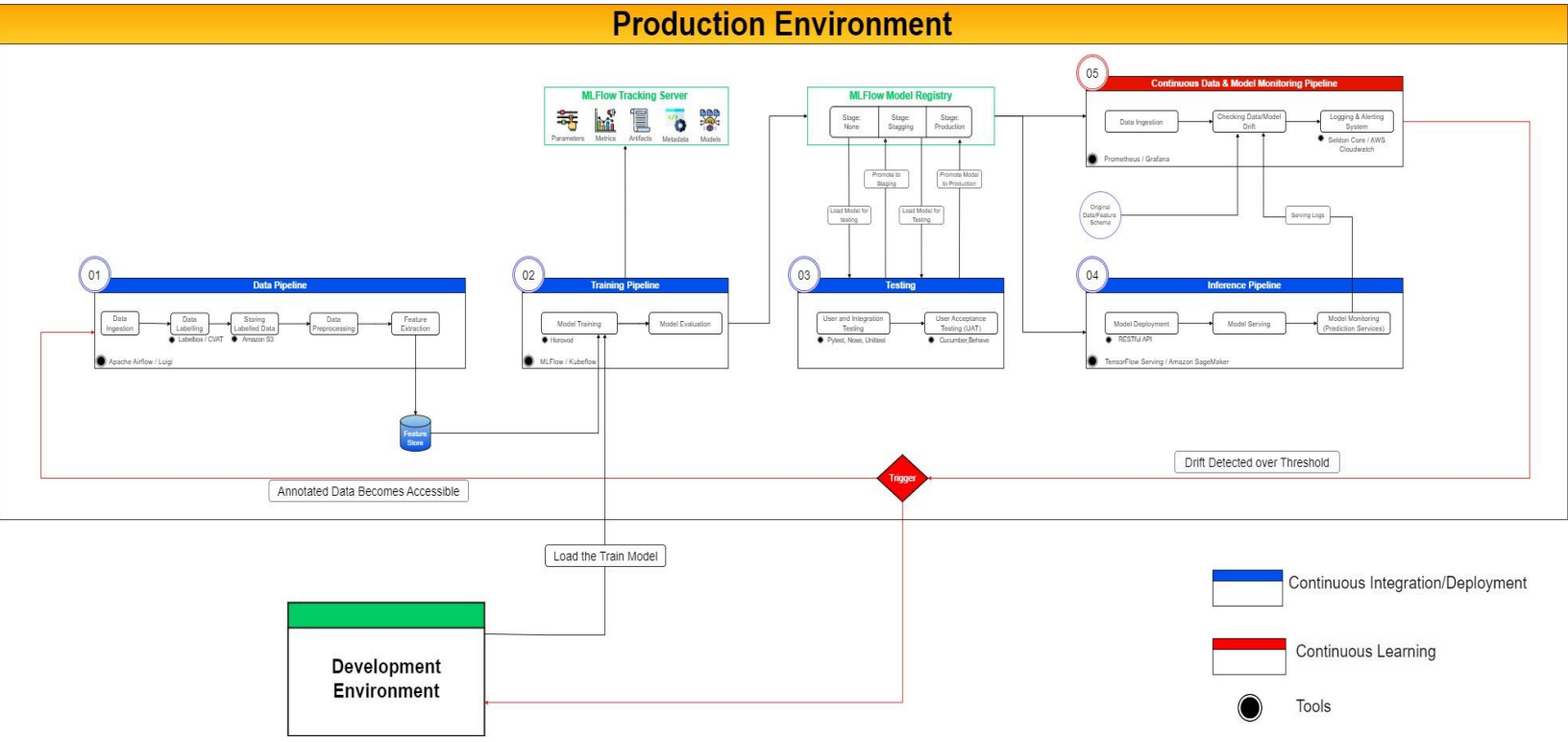
Answer 2: - Building an MLOps system for the used car damage detection problem, as opposed to a simple model, would bring several advantages to the company. These advantages include:

1. **Improved Collaboration:** Developing an MLOps system improves communication between data scientists and DevOps teams, resulting in a more rapid and effective development process.
2. **Improved Model Performance:** MLOps systems include tools for tracking and comparing model performance over time, enabling the business to continuously enhance the model's accuracy and fine-tune its parameters.
3. **Efficient Model Deployment:** MLOps systems offer seamless and automatic model deployment, which allows the model to be deployed to production without the need for manual intervention and with minimal effort.
4. **Model Monitoring:** MLOps solutions offer the capacity to monitor model performance in real-time, ensuring that it continues to operate properly and allowing for fast resolution of any issues that may develop.
5. **Traceability and Audibility:** MLOps solutions include tools for tracking and recording the whole model creation process, including experimentation and deployment, enabling the organization to maintain complete transparency and accountability.
6. **Increased Flexibility:** MLOps platforms are meant to be adaptable, so businesses can easily add extra annotations or train the model on new data as their needs change.
7. **Scalability:** MLOps systems give a scalable, efficient, and automated solution to the problem, allowing the business to process a huge number of photos without requiring manual inspections.
8. **Faster Time-to-Market:** By automating a significant portion of the manual and time-consuming operations involved in the ML pipeline, MLOps technologies enable enterprises to bring ML products to market more quickly.

By Ashwini Abhang IIITB

Q3. System design: You must create an ML system that has the features of a complete production stack, from experiment tracking to automated model deployment and monitoring. For this problem, create an ML system design (diagram)

Answer3: -



To view the image without any distortion use the below link
https://drive.google.com/file/d/1xUgydR7z1-L1AhnN9M11fkqoSxlyNljU/view?usp=share_link

Don’t forget to zoom in!!

Q4. System design: After creating the architecture, please specify your reason for choosing the specific tools you chose for the use case.

Answer4: - The following is a potential approach to developing a machine learning (ML) system for Carsdepo.com to detect damages in used cars:

1. **Data collection and storage:** To construct an object-detection model, a big database of images of used automobiles with annotations indicating the presence or absence of scratches and dents is required. This data can be stored using a cloud-based data storage system, such as ***Amazon S3***, which is easily expandable as the data size grows.
2. **Data preparation and preprocessing:** Before training the model, the data must be organized and preprocessed. This may require dividing the data into training, validation, and testing sets, scaling the images to a uniform size, and transforming the images to a format compatible with deep learning frameworks such as ***TensorFlow*** and ***PyTorch***.
3. **Model Selection and Training:** After data preprocessing, the next step would be to select and train an appropriate object detection model to detect scratches and dents in secondhand cars. ***YOLO*** is a well-known open-source paradigm for object detection (You Only Look Once). ***Faster R-CNN*** is an additional well-known open-source model. The selection of the model will depend on its accuracy, speed, and available resources. Also we can use **Horovad** for large-scale training. We need to specify the architecture of the model and set hyperparameters, such as the learning rate, batch size, and number of epochs. To ensure that the model is generalizing well, we can also use techniques like early stopping, model checkpointing, and cross-validation.
4. **Experiment tracking:** We can use experiment tracking programmes such as ***Weights & Biases***,***Comet.ml, MLFlow or Kubeflow*** to keep track of the different tests we do and the results we acquire. This will enable us to visualize the training procedure, compare several models, and simply duplicate experiments.
5. **Model deployment:** Once the model has been trained, it must be deployed to the production environment. This is possible using deployment solutions such as ***TensorFlow Serving*** or ***AWS SageMaker***, which allow us to provide the model as a REST API.

6. **Monitoring and maintenance:** To guarantee that the model performs effectively in production, we must monitor its performance with tools such as ***TensorBoard*** and ***Datadog*** or ***AWS Cloudwatch***. This will allow us to spot problems early and take corrective action. In addition, we must continually retrain the model to incorporate any newly available annotations and do basic maintenance to keep the system functioning properly.
7. **Version Control:** A version control system, such as ***Git***, may be used to monitor the modifications made to the code.

Q5. Workflow of the solution:
You must specify the steps that should be taken to build such a system end to end.
The steps should mention the tools used in each of the components and how they are connected with one another to solve the problem.
Broadly the workflow should include the following:

- Data and model experimentation
- Automation of data pipeline
- Automation of training pipeline
- Automation of inference pipeline
- Continuous monitoring pipeline

The workflow should also explain the actions to be taken under the following conditions:
After you deployed the model, you noticed that there was a sudden increase in the drift due to the poor lighting in the image taken.

1. What component/pipeline will be triggered if there is any drift detected? What if the drift detected is beyond an acceptable threshold?
2. What component/pipeline will be triggered if you have additional annotated data?

Answer5: - The workflow of the solution for building an automated damage-detection system for Carsdepo.com can be as follows:

1. **Data and model experimentation:**
    - Including both tagged and unlabeled data, a collection of images of used vehicles.
    - Data pre-processing and cleaning, including resizing and converting images to a suitable format.
    - Creating training, validation, and test sets from the data.

- Selecting an appropriate model for object detection, such as YOLO, Faster R-CNN, or RetinaNet.
- Experimenting with several models and hyperparameters to identify the model with the greatest performance.

2. **Automation of data pipeline:**
    - Automate the data pipeline using solutions such as Apache Airflow, Apache NiFi, and AWS Glue.
    - Configure the intake of data from numerous sources, including databases and cloud storage.
    - Automating the preprocessing and cleaning of data, including the scaling and format conversion of images.
    - Creating training, validation, and test sets automatically.
3. **Automation of training pipeline:**
    - Utilize applications such as TensorFlow, PyTorch, or Keras to construct the training pipeline.
    - Automating the training procedure, including the selection of the model and hyperparameters with the highest performance.
    - The automated saving of model weights and outputs.
4. **Automation of inference pipeline:**
    - Utilize tools such as TensorFlow Serving, Flask, or Django to deploy the inference model.
    - Automating the process of producing predictions based on newly collected data, such as images of used vehicles**.**
5. **Continuous monitoring pipeline:**
    - Utilize tools like TensorBoard, Weights & Biases, or Neptune to monitor the model's performance.
    - Continuous monitoring of model performance on the validation and test sets, including accuracy, precision, and recall.
    - If any performance drift is observed, automated alerts and reporting will occur.

Answer for the Drift Conditions due to bad lighting

- **Answer1: -** In the event of a sudden rise in drift due to inadequate lighting in the captured image, the pipeline for continuous monitoring would be triggered. If the drift is observed, the system would investigate the reason for the drift and take necessary action, such as retraining the model with

additional data or modifying the preprocessing procedures. If the drift exceeds a predetermined threshold, the system would automatically retrain the model.

- **Answer2: -** If additional annotated data becomes accessible, the data pipeline's automation would be triggered to incorporate the new data into the training and validation sets. The model would then be retrained on the updated data, and the pipeline for continuous monitoring would monitor the model's performance to ensure it remained within acceptable limits.