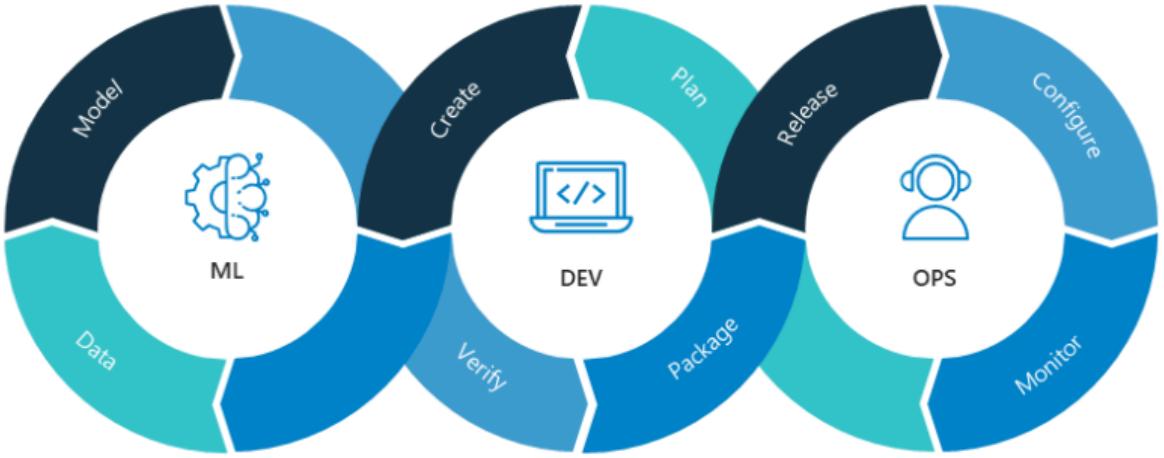
MLOps Requirements



Project Title: easyMoney Data Science Capstone Project
Project Team Members: Gonçalo, Seema, Chukwubuikem
Date: 06/04/2024

INDEX

Task 1 - Analysis

ML Project Stages	Requirements	Constraints	Suggestions (Tech stack, tools, storage, etc)
Data Acquisition	Integration with data sources (databases, APIs, files)	No direct access to the database, limited history partitions and no integration with the new ERP system. Budget limitations.	Backup Data: save data in BigQuery or SnowFlake warehousing platforms; PostgreSQL
	Automated data ingestion pipelines		Apache Beam (Google)
	Data quality monitoring and validation		ReportServer
	Version control for data sets and metadata		Git
Data Preparation	Data preprocessing pipelines		scikit-learn Pipelines
	Feature engineering automation		BigQuery
	Data transformation and normalisation		scikit-learn Transformers
	Version control for preprocessing scripts and configurations		Git

Task 2 - Segmentation

ML Project Stages	Requirements	Constraints	Suggestions (Tech stack, tools, storage, etc)
Model Development	Experiment tracking and management	Budget	MLflow
	Version control for machine learning models and hyperparameters		Git, MLflow
	Automated model training and tuning		AutoML Python libraries such as: PyCaret, H2O AutoML or TPOT
	Integration with ML frameworks and libraries		-
Model Deployment	Containerization of models (Docker)		Docker, Kubernetes
	Automated model deployment pipelines		Kubeflow (build-in Kubernates)
	Monitoring of model performance and health		Prometheus
	Scalability and elasticity of deployed models		Google Cloud Functions (not open source)

Task 3 - Recommendation

ML Project Stages	Requirements	Constraints	Suggestions (Tech stack, tools, storage, etc)
Model Development	Experiment tracking and management	Budget	MLflow
	Version control for machine learning models and hyperparameters		Git, MLflow
	Automated model training and tuning		AutoML Python libraries such as: PyCaret, H2O AutoML or TPOT
	Integration with ML frameworks and libraries		-
Model Deployment	Containerization of models (Docker)		Docker, Kubernetes
	Automated model deployment pipelines		Kubeflow (build-in Kubernates)
	Monitoring of model performance and health		Prometheus
	Scalability and elasticity of deployed models		Google Cloud Functions (not open source)

Task 4 - Monitoring

ML Project Stages	Requirements	Constraints	Suggestions (Tech stack, tools, storage, etc)
Data Acquisition	Real-time monitoring of model predictions	Budget	Prometheus, MLflow
	Anomaly detection and alerting		Elasticsearch
	Drift detection to identify model decay or shifts in data		TensorFlow Model Analysis
	Logging and auditing of model predictions and inputs		Elasticsearch
Model Maintenance	Automated retraining of models		MLflow, Kubeflow
	Feedback loop integration to update models with new data or feedback from users		MLflow
	Version control for model retraining scripts and configurations		Model Evaluation Library (ModEva), MLflow, TensorFlow Model Analysis
	Continuous evaluation of model performance and relevance		Model Evaluation Library (ModEva), MLflow, TensorFlow Model Analysis

Task 5 - Coordination/ Planning

Requirements	Constraints	Suggestions (Tech stack, tools, storage, etc)
Team coordination and task, validation, development planning	Budget	Trello (Not open source)