ML Case Study

Guidelines for the ML Case Study:

About the Project:

Welcome to the realm of hands-on machine learning! In this open-ended ML Case Study, you have the opportunity to explore the vast landscape of machine learning applications across diverse industries. Your mission is to select one project from the list of projects given below in the cheat sheet or propose your own project idea that aligns with your interests and aspirations. This project is your canvas to apply your machine learning skills, experiment with various algorithms, and demonstrate your ability to tackle real-world challenges using data-driven insights. Through this project, you'll showcase your autonomy, creativity, and technical prowess as you contribute to shaping the future with data-driven solutions.

Skills Required:

- Proficiency in Python programming.
- Familiarity with data manipulation libraries (e.g., pandas,numpy).
- Knowledge of data visualisation tools (e.g., matplotlib, seaborn).
- Understanding of machine learning concepts and clustering algorithms (e.g., K-means).
- Ability to analyse and interpret data insights.
- Strong report writing and presentation skills.

Deliverables:

- **Case Study Colab Notebook:** Students should submit a Google colab (.ipynb) showcasing their data analysis process, including loading the dataset, data cleaning, exploration, visualisations, modelling and preliminary insights.

Rubrics for Assessment:

Data Exploration and Preprocessing:

- Correct loading of data and handling of missing values.
- Effective identification and handling of outliers if present.

Feature Engineering:

- Appropriate calculation of metrics and creation of new features.
- Clear explanations for the chosen feature engineering strategies.

Customer Segmentation:

- Proper selection and application of the clustering algorithm.
- Clear justification for the chosen number of clusters.

- Accurate visualisation of customer segments.

Interpretation and Analysis:

- Thorough analysis of customer segments' characteristics.
- Meaningful insights and observations drawn from the analysis.
- Demonstration of critical thinking and depth of understanding.

Recommendations:

- Relevant and actionable recommendations for marketing strategies.
- Alignment of recommendations with the analysis.

Visualisations and Reporting:

- Effective visualisations that enhance the understanding of insights.
- Clear, concise, and well-structured report or presentation.

Problem Statement:

Imagine yourself as a freelance data scientist ready for the next project adventure. Your task is to select a machine learning project from the list provided or propose an original project idea that resonates with you. Your objective is to identify a specific challenge within the chosen industry domain and design a machine-learning solution to address it. Whether you're predicting customer behavior, optimizing processes, or making healthcare more efficient, your project should demonstrate your ability to approach complex problems, preprocess and analyze relevant data, develop and fine-tune models, and interpret results in a meaningful way. Your project will be a testament to your adaptability, curiosity, and aptitude for machine learning.

Execute an end-to-end data science project by following the below steps:

Step 1: Define the Problem Statement

- Understand the industry and categorize the problem type (Supervised, Unsupervised, Semi, etc.).
- Comprehend the business objective and desired outcomes.
- Identify constraints, limitations, computational power, budget, and data availability.
- Determine evaluation metrics for optimization, tracking KPIs, and required testing.
- Assess the model's relevancy to the target audience, focusing on prediction speed.
- Evaluate data availability and necessary features for collection.
- Define the scope of the solution to manage expectations.
- Consider deployment options such as cloud platforms, web apps, websites, or APIs.

Step 2: Data Collection

- Identify reliable sources such as databases, APIs, sensors, or surveys.
- Specify the required data volume for effective analysis.
- Classify data as labeled or unlabeled based on availability.
- Address data quality issues, errors, bias, and consistency.
- Ensure data relevancy to the problem being addressed.
- Account for temporal effects and changes in the data.
- Handle legal and ethical concerns related to data privacy.
- Implement sampling strategies and data privacy techniques.
- Utilize appropriate tools for data collection.
- Implement version control to manage dataset changes.

- Consider continuous data collection for improved accuracy.

Step 3: Data Preprocessing

- Handle missing values using various imputation techniques.
- Address outliers using standard deviation or IQR methods.
- Encode categorical variables using suitable techniques.
- Transform data through standardization, normalization, or other methods.
- Handle imbalanced datasets using techniques like oversampling or undersampling.
- Reduce dimensionality for better computational efficiency.
- Apply techniques to transform data for optimal model performance.

Step 4: Exploratory Data Analysis (EDA)

- Analyze data distribution using summary statistics and visualizations.
- Explore relationships between variables through scatter plots and bar charts.
- Study complex interrelationships using heatmaps and pair plots.
- Identify temporal patterns and trends.
- Visualize categorical data using appropriate charts.
- Use PCA for dimensionality reduction and visualization.
- Perform statistical and hypothesis tests to validate assumptions.
- Visualize complex data types such as text or images.

Step 5: Model Selection, Training & Evaluation

- Split data into training and testing sets.
- Choose suitable algorithms from a library based on the problem.
- Select evaluation metrics aligned with the problem domain.
- Ensure scalability and efficient processing for larger datasets.
- Optimize hyperparameters through techniques like grid search.
- Utilize parallel processing and GPU resources for training.
- Interpret and explain model decisions using tools like SHAP or LIME.
- Address imbalanced data to prevent bias in model performance.
- Consider pre-trained models and transfer learning for enhanced training.
- Implement early stopping to prevent overfitting.
- Save and load models for future use.
- Use experiment logging and versioning tools.
- Integrate the model into data processing pipelines.
- Implement feedback loops for retraining based on updated data.
- Explore automated machine learning (AutoML) for model selection and hyperparameter tuning.

Cheat Sheet to Approach the Problem Statement

End to End Steps For Any Data Science Project

1. **Define the problem statement -** Be Sure What you want to do & Achieve
2. **Data Collection -** Take the dataset and store it.
3. **Data Preprocessing -** Prepare the recipe, before the food.
4. **Exploratory Data Analysis (EDA) -** Get the insights
5. **Model Selection, Training & Evaluation -** Choose, train, and assess a suitable model to solve the defined problem.

Detailed steps of doing a ML project:

1. Define the problem statement- Be Sure What you want to do & Achieve

1.1. Understand the Industry - Type of Problem (Supervised, Unsupervised, Semi , etc)
1.2. Understand the Business Objective - Why this problem & Desired Outcome
1.3. Constraints & Limitations - Computational Power, Budget, Data Availability, Obstacle
1.4. Evaluation Metrics - Optimization Required, KPIs Tracking, Required Testing,
1.5. Relevancy to the target Audience - Model Prediction Usage i.e., Speed of Predict
1.6. Data Availability - Ease of Data Collection, Necessary Features Required
1.7.Scope of the Solution- Define the solution's capabilities to effectively manage expectations while addressing the problem.
1.8. Deployment Considerations - Cloud Platform, Webapp or website integration or just API

ML Business Use Cases - Various Relevant Industries

Manufacturing	Retail	Healthcare & Life Sciences
1. **Predictive Maintenance or Condition Monitoring** - Supervised Classification/Regression - Precision/Recall 2. **Warranty Reserve Estimation** - Supervised Regression - MAE/RMSE 3. **Propensity to Buy** - Supervised Classification - AUC-ROC 4. **Demand Forecasting** - Time Series Forecasting - MAPE/RMSE 5. **Process Optimization** - Supervised Regression/Reinforcement Learning - MAE 6. **Telematics** - Supervised Classification/Regression - AUC-ROC/RMSE 7. **Quality Assurance** - Supervised Classification - Precision/Recall 8. **Supply Chain Optimization** - Supervised Regression/Reinforcement Learning - MAE/Cost Savings 9. **Energy Consumption Optimization** - Time Series Forecasting - RMSE/MAPE 10. **Safety Monitoring** - Supervised Classification - Recall **11. & More**	1. **Customer Segmentation -**Unsupervised Clustering - Silhouette Score/Cohesion and Separation 2. **Sales Forecasting -** Time Series Forecasting - MAPE/RMSE 3. **Inventory Optimization -** Supervised Regression/Reinforcement Learning - MAE/Cost Savings 4. **Churn Prediction -** Supervised Classification - AUC-ROC 5. **Recommendation Systems -** Collaborative/Content-based Filtering - RMSE/Precision@K 6. **Price Optimization -** Supervised Regression/Reinforcement Learning - MAE 7. **Fraud Detection -** Supervised Classification - Precision/Recall/F1-Score 8. **Store Location Analysis -** Supervised/Unsupervised Learning - AUC-ROC/Silhouette Score 9. **Customer Lifetime Value Prediction -** Supervised Regression - MAE/RMSE 10. **Market Basket Analysis -** Association Rule Learning - Lift/Confidence **11. & More.**	1. **Disease Prediction -** Supervised Classification - AUC-ROC/Precision/Recall 2. **Drug Discovery -** Supervised Regression/Classification - AUC-ROC/RMSE 3. **Patient Adherence & Retention -** Supervised Classification - Precision/Recall 4. **Medical Image Analysis -** Supervised Classification - AUC-ROC/Precision/Recall 5. **Genomic Data Analysis -** Supervised/Unsupervised Learning - AUC-ROC/Silhouette Score 6. **Treatment Efficacy Analysis -** Supervised Regression - MAE/RMSE 7. **Hospital Readmission Prediction -** Supervised Classification - AUC-ROC 8. **Clinical Trial Optimization -** Supervised Regression/Reinforcement Learning - MAE/Cost Savings 9. **Outbreak Prediction -** Time Series Forecasting - MAPE/RMSE 10. **Patient Segmentation for Personalized Treatment -** Unsupervised Clustering - Silhouette Score/Cohesion and Separation **11. & More.**

Travel & Hospitality	Banking & Financial Services	Energy, Feedstock & Utilities
1. **Demand Forecasting -** Time Series Forecasting - MAPE/RMSE 2. **Dynamic Pricing -** Supervised Regression - MAE 3. **Customer Segmentation -** Unsupervised Clustering - Silhouette Score/Cohesion and Separation 4. **Recommendation Systems for Personalized Travel Packages -** Collaborative/Content-based Filtering - RMSE/Precision@K 5. **Churn Prediction for Loyalty Programs -** Supervised Classification - AUC-ROC 6. **Sentiment Analysis from Customer Reviews -** Supervised Classification - AUC-ROC/Precision/Recall 7. **Optimal Route Analysis -** Supervised Regression/Reinforcement Learning - MAE/Cost Savings 8. **Fraud Detection in Bookings/Transactions -** Supervised Classification - Precision/Recall/F1-Score 9. **Hotel Room Occupancy Prediction** - Time Series Forecasting/Supervised Regression - MAPE/RMSE 10. **Event-based Promotion Targeting -** Supervised Classification - AUC-ROC/Precision/Recall **11. & More.**	1. **Credit Scoring -** Supervised Classification - AUC-ROC 2. **Fraud Detection -** Supervised Classification - Precision/Recall/F1-Score 3. **Customer Churn Prediction -** Supervised Classification - AUC-ROC 4. **Portfolio Management -** Supervised Regression/Reinforcement Learning - Sharpe Ratio/MAE 5. **Algorithmic Trading -** Time Series Forecasting/Reinforcement Learning - Sharpe Ratio/Profit and Loss 6. **Market Basket Analysis for Banking Products -** Association Rule Learning - Lift/Confidence 7. **Customer Segmentation -** Unsupervised Clustering - Silhouette Score/Cohesion and Separation 8. **Risk Management -** Supervised Regression - MAE/RMSE 9. **Credit Card / Loan Default Prediction -** Supervised Classification - AUC-ROC 10. **Sentiment Analysis for Market Trends -** Supervised Classification - AUC-ROC/Precision/Recall **11. & More.**	1. **Demand Forecasting -** Time Series Forecasting - MAPE/RMSE 2. **Energy Theft Detection -** Supervised Classification - Precision/Recall/F1-Score 3. **Predictive Maintenance for Equipment -** Supervised Classification/Regression - Precision/Recall 4. **Optimization of Energy Grids -** Reinforcement Learning - Cost Savings/MAE 5. **Wind and Solar Energy Prediction -** Time Series Forecasting/Supervised Regression - MAPE/RMSE 6. **Oil and Gas Exploration or Resorvior Prediction -** Supervised Regression/Classification - AUC-ROC/RMSE 7. **Customer Segmentation for Energy Services -** Unsupervised Clustering - Silhouette Score/Cohesion and Separation 8. **Price Forecasting for Commodities -** Time Series Forecasting - MAPE/RMSE 9. **Optimization of Feedstock Blends in Production -** Supervised Regression - MAE 10. **Water Quality Analysis -** Supervised Classification - AUC-ROC/Precision/Recall **11. & More.**

Automotive Sector	Telecom & Internet Services	Agriculture & Food Industry
1. **Predictive Maintenance for Vehicles -** Supervised Classification/Regression - Precision/Recall 2. **Driver Behavior Analysis -** Supervised Classification/Time Series Analysis - AUC-ROC/Precision/Recall 3. **Vehicle Sales Forecasting -** Time Series Forecasting - MAPE/RMSE 4. **Autonomous Vehicle Navigation -** Reinforcement Learning/Deep Learning - Accuracy/Error Rate 5. **Supply Chain Optimization for Parts -** Supervised Regression/Reinforcement Learning - MAE/Cost Savings 6. **Quality Assurance in Manufacturing -** Supervised Classification - Precision/Recall 7. **Vehicle Image Recognition & Analysis -** Supervised Classification/Deep Learning - AUC-ROC/Accuracy 8. **Recommendation Systems for Car Features -** Collaborative/Content-based Filtering - RMSE/Precision@K 9. **Fuel Efficiency Prediction -** Supervised Regression - MAE/RMSE 10. **Crash Analysis & Safety Testing -** Supervised Classification/Regression - AUC-ROC/Precision/Recall **11. & More.**	1. **Churn Prediction -** Supervised Classification - AUC-ROC 2. **Network Traffic Analysis and Prediction -** Time Series Forecasting - MAPE/RMSE 3. **Predictive Maintenance for Network Equipment -** Supervised Classification/Regression - Precision/Recall 4. **Fraud Detection for Telecom Services -** Supervised Classification - Precision/Recall/F1-Score 5. **Optimization of Network Configuration -** Reinforcement Learning - Efficiency/Cost Savings 6. **Customer Segmentation -** Unsupervised Clustering - Silhouette Score/Cohesion and Separation 7. **Recommendation Systems for Value-Added Services -** Collaborative/Content-based Filtering - RMSE/Precision@K 8. **Sentiment Analysis from Customer Feedback -** Supervised Classification - AUC-ROC/Precision/Recall 9. **Bandwidth Allocation and Optimization -** Supervised Regression/Reinforcement Learning - MAE 10. **Location-based Service Enhancements -** Supervised Classification/Regression - AUC-ROC/Precision/Recall **11. & More.**	1. **Crop Yield Prediction -** Supervised Regression - MAE/RMSE 2. **Disease Detection in Crops -** Supervised Classification - Precision/Recall 3. **Precision Agriculture (Soil, Water, Fertilizer Optimization) -** Supervised Regression/Reinforcement Learning - MAE 4. **Food Quality Assurance -** Supervised Classification - Precision/Recall 5. **Supply Chain Optimization for Food Products -** Supervised Regression/Reinforcement Learning - MAE/Cost Savings 6. **Demand Forecasting for Agricultural Products -** Time Series Forecasting - MAPE/RMSE 7. **Animal Behavior and Health Monitoring -** Supervised Classification/Time Series Analysis - AUC-ROC/Precision/Recall 8. **Food Recommendation Systems -** Collaborative/Content-based Filtering - RMSE/Precision@K 9. **Agricultural Drone Path Optimization -** Reinforcement Learning - Efficiency/Cost Savings 10. **Genomic Data Analysis for Crop Breeding -** Supervised/Unsupervised Learning - AUC-ROC/Silhouette Score **11. & More.**

2. Data Collection - It’s not a Capstone Project. It’s real Life. Wake up.

2.1. Source Identification: Determine reliable sources like databases, APIs, sensors, or surveys for obtaining data.
2.2. Data Volume Required: Specify the amount of data needed for effective analysis.
2.3. Data Types: Classify data as labeled or unlabeled, based on its availability.
2.4. Data Quality: Address errors, bias, and maintain consistency within the data.
2.5. Data Relevancy: Ensure collected data is directly relevant to the problem being addressed.
2.6. Temporal Considerations: Account for time-related effects, seasonality, and changes in the data.
2.7. Legal and Ethical Concerns: Address data ethics and privacy policies to ensure compliance.
2.8. Sampling Strategy: Opt for sampling instead of using the entire population.
2.9. Data Privacy: Apply techniques like Z-Score or normalization for handling private/sensitive information.
2.10. Data Collection Tools: Utilize web scrapers and other appropriate tools.
2.11. Data Versioning: Implement version control for managing changes in the dataset.
2.12. Continuous Data Collection: Regularly update the model with fresh data over time for improved accuracy.

3. Data Preprocessing - Prepare the recipe, before the food.

3.1. Handling Missing Values

- Mean/Median/Mode Replacement
- Random Sample Imputation
- Capturing NaN value with a new feature
- End of Distribution Imputation
- Arbitrary Imputation - Similar to Hyperparameter Tuning
- Frequent Categories Imputation
- Creating a Sub-model to predict the missing value
- Deleting Column if missing value > 60%
- Using Algorithms Which Support Missing Values i.e. KNN
- Checking for duplicates

3.2. Handling Outliers

- Using Standard Deviation in Symmetric Curve
- Using IQR in skew-symmetric Curve
- Using Outlier Insensitive Algorithms.i.e. SVM, KNN, Decision Tree

3.3. Categorical Encoding

- Nominal Encoding
- One Hot Encoding
- One Hot Encoding with many categorical - KDD Cup Challenge & Reduce Curse of Dimensionality
- Mean Encoding

- Ordinal Encoding
- Label Encoding
- Target guided Ordinal Encoding = mean encoding +label encoding
- Count (or) Frequency Encoding
- Probability ratio Encoding - Titanic Dataset (Survival & Not Survival)

3.4. Data Transformation

- Standardisation
- Normalisation
- Robust Scaler
- Sum of (median-observation)/IQR
- Box-Cox Transformation - transformation of a non-normal dependent variable into a normal shape - T(y)=(y exp(lambda)-1)/(lambda)
- Gaussian Transformation
- Logarithmic Transformation
- Inverse Transformation
- Square Root Transformation
- Exponential Transformation

3.5. Handling Imbalanced Dataset

- Cross-Validation
- Under Sampling
- Over Sampling
- Synthetic Minority over Sampling Technique (SMOTE)
- Tree-Based Algorithm

3.6. Data Reduction

- Dimensionality Reduction
- Numerosity Reduction - original data<>smaller form

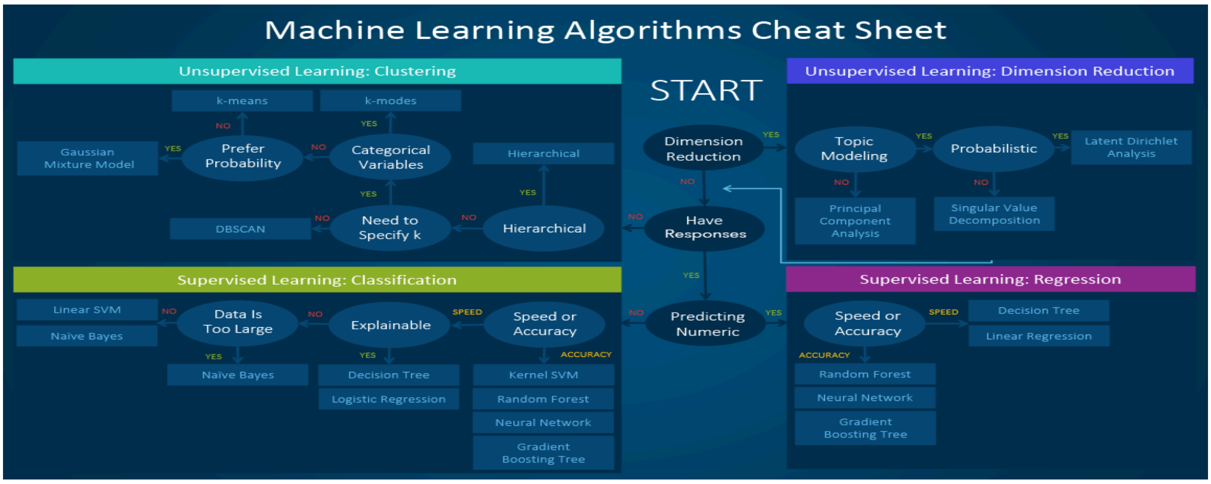
4. Exploratory Data Analysis (EDA) - Uncover Insights, Ignite Understanding.

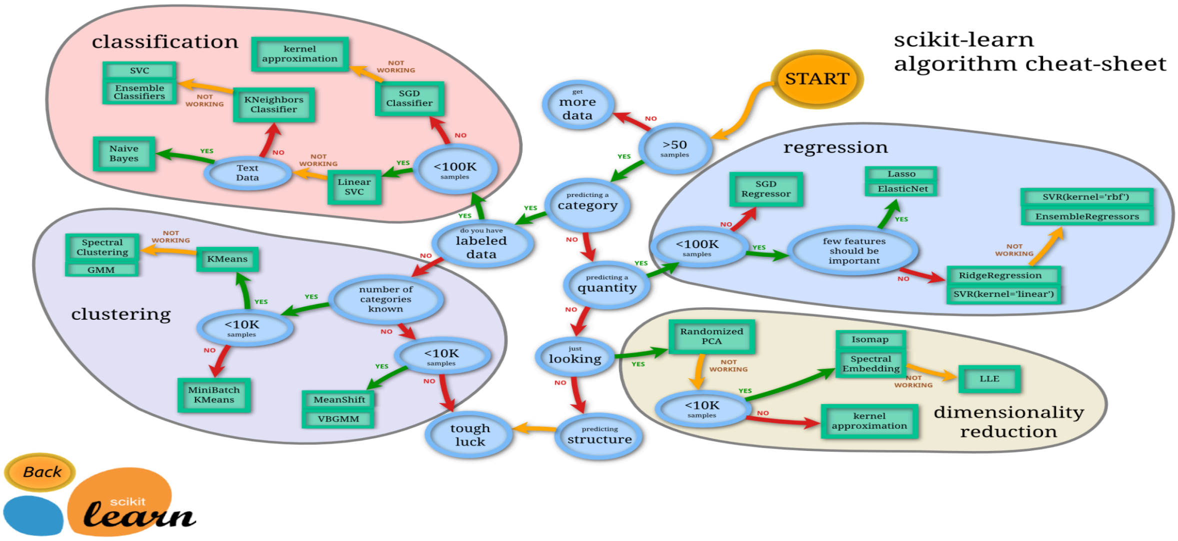
4.1. Distribution Analysis: Utilize summary statistics, histograms, and box plots to understand data distribution.
4.2. Bivariate Analysis: Examine relationships between two variables using scatter plots and bar charts.
4.3. Multivariate Analysis & Feature Relationships: Explore complex interrelationships using tools like heatmaps and pair plots.
4.4. Temporal Analysis: Study data patterns and trends over time.
4.5. Categorical Data Analysis: Visualize categorical data through bar and pie charts**.**
4.6. Dimensionality Reduction Visualization (PCA): Reduce data dimensions and visualize using PCA.

4.7. Statistical & Hypothesis Tests: Apply tests like t-test and chi-square to validate assumptions.
4.8. Complex Data Type Visualization: Visualize text data using word clouds, images with distribution insights.
4.9. Geospatial Analysis: Analyze data in relation to geographical locations.
4.10. Segmentation Analysis: Identify data clusters or segments for further insights.

5. Model Selection, Training & Evaluation - Crafting, Polishing, and Assessing the Winning Model

5.1. Data Splitting: Divide data into various proportions for training and testing i.e 80:20, 70:30, 90:10, 80:10:10, 90:5:5
5.2. Algorithm Library: Select suitable algorithms from a comprehensive library.





5.3. Model Evaluation Metrics: Choose metrics aligned with the specific problem domain. (Healthcare - Recall, Stock Market - Precision) etc.

5.4. Scalability: Ensure the model's capacity to handle larger datasets if required.
5.5. Hyperparameter Tuning: Optimize hyperparameters using techniques like grid search or automated tools.
5.6. Parallelization and GPU Support: Utilize parallel processing and GPU resources for efficient training.
5.7. Model Interpretability & Explainability: Employ tools like SHAP, LIME, and ELI5 to understand model decisions.
5.8. Handling Imbalanced Data: Address imbalances to prevent bias in model performance.
5.9. Transfer Learning and Pre-trained Models: Leverage pre-trained models for faster and enhanced training.
5.10. Early Stopping: Monitor and halt training to prevent overfitting.
5.11. Save and Load Models: Store and retrieve trained models for future use.
5.12. Experiment Logging and Versioning: Use tools like MLflow or Weights & Biases to manage model versions.
5.13. Pipeline Integration: Integrate the model into data processing pipelines.
5.14. Feedback Loops: Implement retraining based on updated data.
5.15. Automated Machine Learning (AutoML): Automatically search for optimal models and hyperparameters.