***Machine Learning Lecture 3***

Overview of an End-to-End ML Project

* An ML project is a comprehensive process from understanding the problem to deploying and monitoring the model. The process includes planning , data acquisition, exploration, preparation, model training, fine tuning and finally launching the system

Look the Big Picture

* Problem Framing: Define the business objective, understand current solutions and decide how the ML model’s output will be used.
* Task identification: Determine the type of ML task (regression, classification, clustering, etc…) and the level of supervision (supervised, unsupervised, online vs batch).
* Performance Measures: Choose suitable metrics (e.g, RMSE and MAE for regression, accuracy or confusion matrices for classification, silhouette score or clustering).

Get the Data

* Data Acquisition: Gather data from various sources (CSV files, databases, compressed files).
* Initial Data Inspection: Examine the structure of the dataset, noting capped values, different scales and potential issues like heavy tails
* Test Set Creation: Prevent overfitting by splitting data into training and test sets. Consider random sampling versus stratified sampling (e.g. Keeping income distributions consistent) and use a fixed random seed (commonly 42) for reproducibility.

Explore and Visualize the Data

* Visualization: Use geographical plots, scatter matrices and correlation matrices to understand relationships between features and the target.
* Correlation Analysis: Compute Pearson’s r to identify linear relationships and experiment with feature combinations to uncover new insights.

Prepare the Data from ML Algorithms

* Data cleaning: Address missing values (e.g: using imputation with median or mean) and remove or correct errors
* Categorical Transformation: Convert categorical data using techniques like vocabulary mapping or one-hot encoding
* Numeric Transformation: Normalize numeric features via min-max scaling or z-score standardization; use transformation (e.g., log transformation or bucketisation) to handle skewed distributions.
* Pipelines and Custom Transformers: Build reproducible data transformation pipelines using tools like Scikit-Learns Pipeline and consider writing custom transformers for repeated use.

Select and Train a Model

* Initial Model Selection: Start with a baseline model (such as linear regression ) to gauge performance on the training set.
* Cross-Validation: Use techniques like k-fold cross-validation to access model performance reliably and to understand error variability.
* Underfitting vs Overfitting: Monitor model performance to ensure the model is complex enough to learn patterns without overfitting to the training data.

Fine Tune your Model

* Hyperparameter Tuning: Use grid search or randomized search (e.g., with GridSearchCV) to experiment with different hyperparameter combinations and analyze model errors to refine the model.
* Test Set Evaluation: After fine-tuning, evaluate the best model on the separate test set to measure its true generalization performance.

Launch, Monitor and Maintain

* Deployment: Save your model (e.g., with joblib) and deploy it as a web service, embedded in a website or on a cloud platform.
* Monitoring: Continuously track the models performance in production to detect issues like model rot.
* Maintenance: Regularly update datasets retrain the model, and adjust the system as new data arrives to ensure sustained performance

***Quiz Questions and Answers***

1. Q: What are the seven main stages of an end to end ML project?  
   A: The seven stages are:  
   1) Look at the Big Picture  
   2) Get the data  
   3) Explore and Visualize the Data  
   4) Prepare the Data for ML Algorithms   
   5) Select and Train a model   
   6) Fine-Tune your Model  
   7) Launch, Monitor and Maintain
2. Q: How do you frame the problem in the initial stage of an ML project?  
   A: Framing the problem involves defining the business objective understanding what the client aims to achieve, analyzing current solutions and determining the type of ML task (e.g., regression, classification) along with the appropriate performance measures.
3. Q: Why is creating a separate test set important in an ML project?  
   A: A separate test set ensures that you have unbiased data to evaluate the model’s generalization performance. It prevents data snooping and overfitting by keeping a portion of data unseen during training and tuning.
4. Q: What are some common methods for handling missing data during the data preparation stage?  
   A: Common methods include ignoring instances with missing values, dropping the feature altogether or using imputation techniques such as replacing missing values with the median or mean of the feature using tools like Scikit-Learns’s Simplelmputer
5. Q: Explain the difference between random sampling and stratified sampling for creating test sets.  
   A: Random sampling splits the dataset without considering any underlying distributions which may lead to sampling bias if the dataset is small or imbalanced. Stratified sampling divides the data into homogeneous subgroups (strata) and samples from each subgroup proportionally ensuring that important attributes (like median income) maintain their distribution in both training and test sets.
6. Q: What is the purpose of cross-validation in model training?  
   A: Cross-validation, such as k-fold cross-validation, helps estimate the model’s performance more reliably by splitting the training data into multiple folds. It provides not only the mean error but also the variability (standard deviation) of the error, reducing the risk of overfitting to a single validation set.
7. Q: How does grid search contribute to fine-tuning a model?  
   A: Grid search automates the process of hyperparameter tuning by exhaustively searching through a specified set of hyperparameter values. It uses cross-validation to evaluate the performance of each combination, enabling you to select combination that yield the best performance.
8. Q: What are the common normalization techniques used for numeric data transformation?  
   A: The common normalization techniques are:  
    Min-Max Scaling: Rescales the Features to a specific range (e.g., 0 to 1 or -1 to 1)  
    Z-Score Standardization: Transforms the data so that it has a mean of 0 and a standard deviation of 1. Additionally, log transformation or bucketisation may be used for handling skewed distributions.
9. Q: What techniques are available to transform categorical attributes for ML models?  
   A: Categorical attributes can be transformed using:  
    Vocabulary Mapping: Assigning a unique number to each category, which works best for ordered categories.  
    One-Hot Encoding: Converting each category into binary vectors (1 for the presence and 0 for the absence of a category), especially useful for unordered categories.
10. Q: Once a model is deployed, what strategies are used to ensure it continues performing well over time?  
    A: After deployment its crucial to monitor the models live performance by tracking key metrics and watching for model rot. Maintenance strategies include regular retraining with fresh data, automating retraining pipelines and setting up alerts for anomalies or shifts in data distributions.