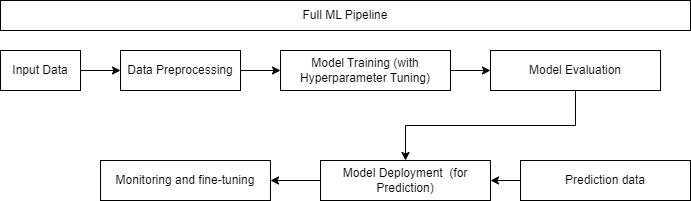
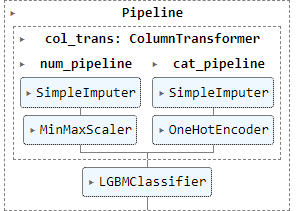
# Solution Design for the ML Pipeline



* Obtain the input data (most likely in .csv) and run data preprocessing before model training process:
  + Numerical features – impute null values with mean values and apply min-max scaler
  + Categorical features – impute null values with most frequent values (mode) and apply one-hot encoding
* Model Training process with built-in hyperparameter process using GridSearchCV
  + Parameter grid is used to put in different values for each parameter
  + GridSearchCV will choose the best parameter and use it to fit the model
  + Sample of the proposed pipeline structure:



* Model Evaluation will be done with Classification report to evaluate the model performance against each class, this case would be against class 0 and class 1 (binary classification)
  + The optimal model would be high recall, high precision and high accuracy in order to get the predictions right for class 0 and class 1 (to get the low-risk and high-risk loan prediction right)
* Model Deployment: in this demo the best model is exported into a joblib file and placed under the “models” directory.
  + A function to get the latest models is included in the pipeline to load the best available models from the directory based on modified date.
  + The model file naming (.joblib) have also included the model score and the created date.
  + The deployed model could use be use for prediction when there is new data comes into the pipeline (covered in the demo)
* Monitoring and fine-tuning: the model performance should be monitored and apply fine-tuning if needed when possible feature drifts / model’s metrics starts to decrease and have a drop in the accuracy.

**Folder structure:**

***ml\_pipeline***  
 |\_ ***data*** : this folder contains the data needed to train/predict  
 |\_ ***documentations*** : this folder contains the documentation related to the pipeline design  
 |\_ ***models*** : this folder consists the trained model pipeline to be used in production(old version and new version of the model pipeline shall be stored here)  
 |\_ ***requirements.txt*** : to be used to install the required packages to run the machine learning pipeline  
 |\_ ***ML\_full\_pipeline.ipynb*** : the jupyter notebook file consisting of the full machine learning pipeline