



mlangles Predictive AI

Bank Loan Defaulter Prediction







Objective of the use case

The objective of this dataset is to build a predictive model that can identify individuals who are likely to default on their loans based on various features such as funded amount, loan balance, location, etc. The primary goal is to assist banks in mitigating risks associated with loan defaults by enabling them to take preventative measures, thus reducing financial losses and stabilizing economic growth.





Explanation of the use case

The dataset for this hackathon comprises two separate sets: a training dataset and a testing dataset. Here is an overview of each, along with what you can expect from them:

Training Dataset

Size: The training dataset contains 67,463 rows, each representing a loan instance or a customer. There are 35 columns, which include various attributes that describe the loan, the customer, and other related data points.

Purpose: The training dataset is used to develop and fine-tune a machine learning model. It includes information on whether a customer has defaulted on a loan, which serves as the target variable for model training.

Attributes: The columns in this dataset may contain information such as:

- Loan details like funded amount, loan balance, interest rate, and loan term.
- Customer details like location, employment status, and income level.
- Other factors that might impact loan repayment, such as credit score or previous default history.

Target Variable: This is the column indicating whether the customer has defaulted on their loan. The target variable is what the model aims to predict.

Testing Dataset

Size: The testing dataset has 28,913 rows with 34 columns. This dataset is smaller in size and lacks the target variable (whether a loan was defaulted).

Purpose: The testing dataset is used to evaluate the performance of the trained machine learning model. The goal is to ensure the model generalizes well to new data, avoiding overfitting or underfitting.

Attributes: The columns in the testing dataset are similar to those in the training dataset, except they don't include the target variable (loan default). This allows for an unbiased assessment of the model's accuracy.





Working of the use case

Step 1: Data Engineerining and Pipeline Creation



Install Dependencies: Essential packages and libraries have been installed.

Data Extraction: Data extraction involves gathering raw data from the CSV file.

Data Analysis: Once the raw data is collected, the next step is to analyse it to gain insights and understand its characteristics. Data analysis involves examining the structure, patterns, and relationships within the data. This step helps in identifying trends, outliers, or any anomalies present in the dataset. Techniques such as descriptive statistics, correlation analysis, and data profiling are commonly used during this stage.

Data Preprocessing: After analyzing the data, it's common to encounter inconsistencies, missing values, or errors that need to be addressed. Data cleaning involves preprocessing the data to ensure its quality and reliability. This may include tasks such as inputting missing values, removing duplicates, standardizing formats, and handling outliers. The goal is to prepare the data for further analysis and modeling. Data Visualization: Data visualization is a powerful tool for exploring and communicating insights from the data. Visualizations such as box plots, histograms, and heat maps are used to represent different aspects of the data distribution, relationships, and trends. Box plots are useful for visualizing the distribution of numerical data and detecting outliers. Histograms provide a graphical representation of the frequency distribution of continuous variables. Heat maps are effective for visualizing the correlation between variables in a tabular dataset.









Box Plots: They are commonly used in exploratory data analysis to compare distributions across different groups or categories, identify outliers, and assess the variability within a dataset.

Dist Plots:

 $\overset{\textcircled{0}}{\bowtie}$ Identifying data patterns and anomalies.

Understanding the distribution before modeling or statistical analysis.

© Comparing distributions between different groups or categories.

 Assessing assumptions for statistical tests (like normality).





Feature Engineering: This step aims to extract relevant information from the data and represent it in a format that is suitable for modelling. Feature engineering techniques include encoding categorical variables, scaling numerical features, creating interaction terms, and extracting domain-specific features. The goal is to enhance the predictive power of the model by providing it with informative and discriminative features.

| cho n | nlangles i ML | .Ops ML | angles Demo | User 1 W | orkspace / All | Projects/ B | ank Loan De | efaulter Pipeline/ B | Bank Loan | Defaulter #7 | | | | | | | | | | ۵ 😩 |) |
|--------------------|----------------|------------------|------------------------------|-----------------------------|----------------------------------|-------------|---------------|------------------------|-----------------------|-----------------------------------|----------------------------|-----------------------------|-----------------|---------------------------------|----------------------------------|------------------------|-------------------|---------------------------|-------------------------------|-------------------------------|-----|
| A Home | PIPELINES | EXPE | RIMENT TRAC | KING | | | | | | | | | | | | | | | | | |
| Apyter Notebook | | | | 2. Loadi Time 1274 ms | ing The Data Status SUCCES | 5 | \rightarrow | | 4. Cla Tim 1708 | eansing The Da Stat ms SUCC | a D US ESS | <u> </u> | | 6. Feature I Time 4714 ms | Engineering Status SUCCESS | - | | | | | |
| | LOGS | DATA V | ISUALIZATION | | | | | | | | | | | | | | | | | | |
| Experiments | LOAN AMOUNT | FUNDED AMOUNT | FUNDED AMOUNT INVESTOR | TERM | INTEREST RATE | GRADE | SUB GRADE | VERIFICATION STATUS | LOAN TITLE | DEBIT TO INCOME | DELINQUENCY - TWO YEARS | INQUIRES - SIX MONTHS | OPEN ACCOUNT | PUBLIC RECORD | REVOLVING BALANCE | REVOLVING UTILITIES | TOTAL ACCOUNTS | INITIAL LIST STATUS | TOTAL RECEIVED INTEREST | TOTAL RECEIVED LATE FEE | REC |
| erving | 30458 | 9026 | 9378 | 59 | 8.95916 | | | | 42 | 36.8882 | | | 21 | | 670 | 59.19844 | 22 | | 3136.79 | 0.03920 | 2.1 |
| f ø | 18334 | 11635 | 9943 | 58 | 12.8591 | | | | 49 | 22.4281 | | | | | 2834 | 64.17055 | 20 | | 2087.33 | 0.08828 | 1.3 |
| Model Hub | 6251 | 5302 | 14207 | 59 | 5.94416 | | | | 49 | 7.84397 | | | 10 | | 7278 | 78.46078 | 16 | | 163.836 | 0.08451 | 159 |
| Monitoring | 7007 | 21087 | 32962 | 58 | 13.7535 | | | | 49 | 37.2228 | | | | | 4980 | 65.30700 | 26 | | 1618.97 | 0.05315 | 6.0 |
| | 15981 | 15068 | 25006 | 59 | 11.0520 | | 24 | | 38 | 15.2320 | | | | | 8115 | 52.14875 | 18 | | 2109.78 | 0.02049 | 0.1 |
| | 19728 | 4279 | 6150 | 59 | 15.0672 | | | | 38 | 22.8888 | | | 14 | | 2697 | 55.98992 | 24 | | 429.576 | 0.01005 | 3.6 |
| | 20858 | 20848 | 26666 | 59 | 17.0620 | | | | 49 | 21.8530 | | | 24 | | 3977 | 23.16245 | 29 | | 3610.24 | 0.12254 | 5.3 |
| | 19558 | 7577 | 8075 | 59 | 16.1217 | | | | 38 | 5.38608 | | | 19 | | 9891 | 85.70840 | 16 | | 2774.73 | 0.02099 | 5.3 |
| | 34985 | 21578 | 15729 | 58 | 19.1846 | | | | 49 | 35.5128 | | | 14 | | 4710 | 95.15842 | 18 | | 741.906 | 0.06574 | 8.7 |
| | 30993 | 7756 | 22010 | 58 | 10.1233 | | | | 49 | 25.9784 | | | | | 5760 | 68.95204 | | | 1282.09 | 0.05567 | 0.6 |
| | | | | | | | | | | < | 1 of 122 | 44 | | | | | | | | | |

Data Versioning:

- Various processed data versions can be generated through different transformations applied to the same raw dataset, such as deleting columns or applying various transformations on specific columns.
- Throughout the data pipeline, diverse transformations can be executed at each iteration. Consequently, the resulting data at the pipeline's end is systematically versioned.
- Given that each version of the final data is distinct, models trained on these different versions will exhibit varying behaviors.







STEP 2: Experriment Tracking-Modelling with Hyperparameter Optimization

After the data has been prepared and cleaned, the subsequent step involves training a model using this refined dataset. Since the problem at hand is a classification task, there are several models suitable for this purpose. Common options include the AdaBoost Classifier and K-Neighbors Classifier.

LinearSVC (Support Vector Classification): The goal is to maximize the margin between classes, which is the distance between the hyperplane and the closest data points from each class (support vectors). It uses different techniques, such as regularization, to avoid overfitting and to generalize better.

Logistic Regression: It works by modeling the log-odds of a class as a linear combination of the input features. The logistic function maps these log-odds to probabilities between 0 and 1, allowing for probabilistic interpretation.



| c}₀ m | langles IMLOps MLangles Demo User 1 Workspace / Projects/ Bank Loan Defa | ulter/ Experiment Tracking | | | ŵ | |
|---|--|-------------------------------|-----------------|--------------------------------|---|--|
| Ame | PIPELINES EXPERIMENT TRACKING | | | | | |
| (interpretation) Jupyter Notebook | Run Name Run_6_BLD_LogisticRegression_LinearSVC | Learning Method Supervised | | Problem Type Classification | | |
| Projects | | | | | | |
| Pipelines | Instance Type | Data Version | | Target Variable | | |
| ₽ | C6a.8xlarge 🗸 | Clean_artifact V7 | | Loan Status | | |
| Experiments Serving Model Hub | SELECT THE ALGORITHM | | | | | |
| ک Monitoring | AdaBoostClassifier BernoulliNB | DecisionTreeClassifier | DummyClassifier | | | |
| | GradientBoostingClassifier KNeighborsClassifier | LinearDiscriminantAnalysis | LinearSVC | LogisticRegression | | |
| | QuadraticDiscriminantAnalysis RandomForestClassifier | RidgeClassifier | | | | |
| | HYPERPARAMETER OPTIMIZATION | | | | | |
| | Optimization Techniques | Number of Trails | | | | |
| | Optuna 👻 | | | | | |
| | | | | | | |

Additionally, to enhance model performance, a hyperparameter optimization technique called Optuna is employed. Optuna automates the process of tuning hyperparameters, such as learning rate or tree depth, to find the optimal configuration that maximizes model performance. This approach ensures that the model is fine-tuned to achieve the best possible results on the given dataset, improving its accuracy and predictive power.





After generating the run, a range of information can be extracted, such as visualizations of hyperparameters for the optimal algorithm, parameters employed during training, and metrics along with artifacts. These observations offer crucial insights into the model's performance and characteristics, facilitating comprehension of its efficacy and areas where enhancements can be made.

| ~~~~ | O mlangles IMLOps MLangles Demo User 1 Workspace / Projects / Bank Loan Defaulter Experiment Tracking | | | | | | | | | | |
|---------------------|---|----------|----------|------------|-------------------------|----------|---------------|--|--|--|--|
| A Home | PIPELINES EXPERIMENT TRACKING | | | | | | Go to Serving | | | | |
| Lapyter Notebook | | | | | + New Run 🗶 Run Configu | ration | 1 Delete | | | | |
| ۲ | RUN ID | RUN NAME | STATUS 🔻 | CREATED BY | START TIME | END TIME | | | | | |
| Projects | | | Success | | | | | | | | |
| Pipelines | | | Success | | | | | | | | |
| Experiments | | | Success | | | | | | | | |
| | | | Success | | | | | | | | |
| Serving | | | Success | | | | | | | | |
| Model Hub | | | Success | | | | | | | | |
| Monitoring | | | | | | | | | | | |

After a run is created there is a console where we can find the hyper parameters for each trial that has been taken.

| - | Se mlangles IMLOps MLangles Demo User 1 Workspace / Projects / Bank Loan Defaulter Experiment Tracking 🔞 📳 | | | | | | | | | |
|----------------------------------|--|--|---------------------------------------|--|-----------|-----------------------------|--|--|--|--|
| A Home | PIPELINES EXPERIMENT TRACKING | | | | | Go to Serving | | | | |
| Read and () appending () | Experiments List Search Experiment Q | ★ Run Name: Run_5_BLD_AdaBoostClassifier_RandomFore fier | : :stClassifier_DecisionTreeClassi | Run ID: 8b01636fe901495b8be64c0b64c5357e | 📫 Created | I AT: 5/1/2024, 10:40:28 AM | | | | |
| Projects | Exp Name: Run_6_BLD_LogisticRegression Running | Console | Visualization | Parameters | Metrics | Artifacts | | | | |
| Fipelines | | | | | | | | | | |
| Experiments | Exp Name: Run_5_BLD_AdaBoostClassifier A Success Bb01636/e901495b8be64c0b54c5357e | | | | | | | | | |
| Serving | | | | | | | | | | |
| Model Hub | Exp Name: Run 4 - ExtraTrees, DecisionTre à Success a517dsa4e801s40x0b71478152b6x4153 | | | | | | | | | |
| Monitoring | | | | | | | | | | |
| | Exp Name: Run 3 - DecisionTree, ExtraTrees | | | | | | | | | |
| | | | | | | | | | | |
| | Exp Name: Run 2 - ExtraTrees, KNaighbors, | | | | | | | | | |
| | 😸 Exp Name: Run 1 - All Algorithms 🎍 Success | | | | | | | | | |





HYPERPARAMETER VISUALIZATION:

Visual representations of hyperparameters illustrate the influence of various parameter settings on model performance, aiding in the identification of optimal configurations.

The different types of plots that represent the values of the parameters for each trail are as follows:

- Optimization History Plot: It works by modeling the log-odds of a class as a linear combination of the input features. The logistic function maps these log-odds to probabilities between 0 and 1, allowing for probabilistic interpretation.
- Slice Plot: A Slice Plot depicts the correlation between two hyperparameters while keeping the values of other hyperparameters constant. This visualization enables the exploration of interactions among hyperparameters and their impact on model performance, aiding in the discovery of optimal parameter combinations.
- Hyperparameter Importances Plot: The Hyperparameter Importances Plot ranks the significance of hyperparameters according to their impact on model performance. This visualization assists in identifying the most influential hyperparameters, informing subsequent optimization endeavors or strategies for feature selection.
- Parallel Coordinate Plot: The Parallel Coordinate Plot represents high-dimensional hyperparameter spaces by depicting each hyperparameter as a vertical axis and each point in the plot as a hyperparameter configuration. Lines connecting points illustrate







PARAMETERS

Parameters has the best model among all the models that were selected.

| co m | Smlangles IMLOps MLangles Demo User 1 Workspace / Projects / Bank Loan Defaulter Experiment Tracking | | | | | | | | | |
|--------------------------------------|--|------------------------------|------------------------|---|-----------------------|-------------------|--|--|--|--|
| Home | PIPELINES EXPERIMENT TRACKING | | | | | Go to Serving | | | | |
| (a) Jupyter | Experiments List Search Experiment Q | オ Run Name: Run_6_BLD_Logist | icRegression_LinearSVC | II Run ID: 68524b1dd35e408f86caa3b0842c211: | 2 🛗 Created AT: 5/1/2 | 2024, 11:06:28 AM | | | | |
| Notebook | | Console | Visualization | Parameters | Metrics | Artifacts | | | | |
| Projects | Exp Name: Run_6_BLD_LogisticRegression & Success 68524b1dd35e408/96caa3b0842c2112 | NAME | VALUE | | | | | | | |
| Pipelines | | | | | | | | | | |
| Experiments Generations | Exp Name: Run, <u>5</u> , <u>BLD</u> , <u>AdaBoostClassifier</u> . Bc0163064001499:bitw64/cb64c5357e Created by Millangles demo user 1 51/2024, 10.4028 AM | | | | | | | | | |
| Model Hub Model Hub Monitoring | Exp Name: Run 4 - ExtraTrees, DecisionTre & Success a512/date8014400087147815256x4153 Created by MLungles demo user 1 2/6/2024.1149.15 PM | | | | | | | | | |
| | Exp Name: Run 3 - DecisionTree, ExtraTrees Ascess 30cb1x/33db4/220x898xb/7260-5511 Created by Mungles domo user 1 2x6/2024.1138.27 PM | | | | | | | | | |
| | Exp Name: Run 2 - ExtraTrees, KNeighbors, Second 0x:2500x18044094ssx27x7bc7474c28 Created by Mungles domo user 1 2007024, 700.14 FM Exp Name: Run 1 - All Algorithms <u>+ Second</u> | | | | | | | | | |

METRICS

Metrics such as accuracy offer insights into the model's performance. Accuracy measures overall correctness, precision assesses the accuracy of positive predictions, recall emphasizes capturing all positive instances, and the F1 score balances precision and recall. These metrics collectively aid in evaluating the effectiveness and robustness of classification models.

| | 🗞 mlangles IMLOps MLangles Demo User 1 Workspace / Projects / Bank Loan Defaulter Experiment Tracking 🛞 🙎 | | | | | | | | | | |
|---------------------|---|--|--|---------------------------------------|---------------------------|---------------|--|--|--|--|--|
| Home | PIPELINES EXPERIMENT TRACKING | | | | | Go to Serving | | | | | |
| Jupyter Notebook | Experiments List Search Experiment Q | ★ Run Na Run_5_BLD_AdaBoostClassifier_Randomf fier | me: ForestClassifier_DecisionTreeClassi | Run ID: 8b01636fe901495b8be64c0b64c53 | AT: 5/1/2024, 10:40:28 AM | | | | | | |
| Projects | Exp Name: Run_6_BLD_LogisticRegression @ Running | Console | Visualization | Parameters | Metrics | Artifacts | | | | | |
| | | NAME | VALUE | | | | | | | | |
| | 🔹 Exp Name: Run_5_BLD_AdaBoostClassifier 🍐 Success | | | | | | | | | | |
| | - 8b01636fe901495b8be64c0b64c5357e | | | | | | | | | | |
| Serving | | | | | | | | | | | |
| Model Hub | ★ Exp Name: Run 4 - Extra Trees, Decision Tre ▲ Surves | | | | | | | | | | |
| ų, | a517da8e801d40c0b71478152b6c4153 | | | | | | | | | | |
| Monitoring | | | | | | | | | | | |
| | Exp Name: Run 3 - DecisionTree, ExtraTrees 🎍 Success 30xb1x03xb444250x984b477280x5511 | | | | | | | | | | |
| | | | | | | | | | | | |
| | Exp Name: Run 2 - ExtraTrees, KNeighbors, & Success 0xc250a1804409esa27a7bc74f2ic28 | | | | | | | | | | |
| | | | | | | | | | | | |
| | 📄 💩 Exp Name: Run 1 - All Algorithms 🎍 Success | | | | | | | | | | |





DOWNLOADING ARTIFACTS

Artifacts is the model that is trained, and we can download the model for further inference.

| ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ | O mlangles IMLOps MLangles Demo User 1 Workspace / Projects / Bank Loan Defaulter Experiment Tracking | | | | | | | | | | |
|--|---|--|-------------------------------|--|-----------|-----------------------------|--|--|--|--|--|
| Home | PIPELINES EXPERIMENT TRACKING | | | | | Go to Serving | | | | | |
| | Experiments List Search Experiment Q | ★ Run Name: Run_5_BLD_AdaBoostClassifier_RandomForest fier | Classifier_DecisionTreeClassi | Run ID: 8b01636fe901495b8be64c0b64c5357e | 🛗 Created | I AT: 5/1/2024, 10:40:28 AM | | | | | |
| Projects | Exp Name: Run_6_BLD_LogisticRegression | Console | Visualization | Parameters | Metrics | Artifacts | | | | | |
| Pipelines | | ADABOOSTCLASS | FIER 🛓 | | | Model Hub | | | | | |
| Experiments | Exp Name: Run_5_BLD_AdaBoostClassifier & Success Bio163649014992886440844:5357e | DecisionTreeCLA RANDOMFORESTC | LASSIFIER | | | | | | | | |
| Serving | | | | | | | | | | | |
| Model Hub | Exp Name: Run 4 - ExtraTrees, DecisionTre a517da8e801d40c0b71478152b6c4153 | | | | | | | | | | |
| Monitoring | | | | | | | | | | | |
| | Exp Name: Run 3 - DecisionTree, ExtraTrees 30:b1a038b4/42b0a984b47/260:5511 | | | | | | | | | | |
| | | | | | | | | | | | |
| | Exp Name: Run 2 - ExtraTrees, KNeighbors, & Success Oec25/0a180d409daa27a7bc74f2fc28 | | | | | | | | | | |
| | | | | | | | | | | | |
| | 📚 Exp Name: Run 1 - All Algorithms 👍 Success | | | | | | | | | | |

Model Hub: The best model can be pushed to a model hub where it is deployed and exposed as a REST API endpoint. This endpoint allows applications to send new data to the model for making predictions. By integrating the endpoint into applications, users can easily leverage the model's predictive capabilities in their workflows, enabling real-time decision-making based on the model's insights.







Step 3: Serving

In a serving context, we provide values for each column of data and use this information to predict whether a person has diabetes or not. The model learns from historical data during training and uses this knowledge to make predictions on new data.

| ~~~ n | nlangles I MLOps MLangle | es Demo User 1 Workspace / S | erving / Online Serving | | | | | | | 2 |
|--|--------------------------|------------------------------|-------------------------|------------------------|----------------------------------|-------------|------------------------|-------------|--|---|
| Home | ONLINE PREDICTIONS | | | | | | | | | |
| (in the second s | Select Project Name | | | Select Experiment Name | | | Select Model | | | |
| Projects | Bank Loan Defaulter | | | Run_5_BLD_AdaBoostCla | ssifier_RandomForestClassifier_D | Decision Tr | RandomForestClassifier | | | |
| | Enter the inputs | | | | | | | | | |
| | | | | | | | | | | |
| Experiments | Loan Amount | 10000 | Funded Amount | 32236 | Funded Amount Invest or | 12329.36 | Term | 59 | | |
| CE Serving | | | | | | | | | | |
| Model Hub | Interest Rate | 11.10500686 | Grade | | Sub Grade | | Verification Status | | | |
| ئ Monitoring | Loan Title | | Debit to Income | 16.28475781 | Delinquency - two yea rs | | Inquires - six months | | | |
| | | | | | | | | | | |
| | Open Account | 13 | Public Record | | Revolving Balance | 24246 | Revolving Utilities | 74.93255103 | | |
| | Total Accounts | | Initial List Status | | Total Received Interest | 2929.646315 | Total Received Late Fe | 0.10205519 | | |
| | | | | | | | e | | | |
| | Predict Result | 0 | | | | | | | | |
| | | | | | | | | | | |

| | nlangle | s IMLOps Jennyfer Labadie Workspace / Model Hub | | | | | | ۵ 😩 |
|---------------------|---------|---|------------|------------|---------|--------|-----------------|-----|
| Ame Home | Мо | del Hub | | | | | | |
| Jupyter Notebook | | | | | | | T, Clear Filter | |
| ۲ | | MODEL NAME | CREATED BY | CREATED AT | VERSION | STATUS | | |
| Projects | | | | | | | | |
| Pipelines | | | | | | | | |
| ₽ | | | | | | | | |
| Experiments | | | | | | | | |
| Serving | | | | | | | | |
| Model Hub | | | | | | | | |
| من | | | | | | | | |
| Monitoring | | | | | | | | |
| | | | | | | | | |
| | | | | | | | | |
| | | | | | | | | |
| | | | | | | | | |
| | | | | | | | | |
| | | | | | | | | |
| | | | | | | | | |





Step 4: Monitoring

Monitoring for data drift involves regularly assessing the distribution of patient data over time and updating machine learning models accordingly. This ensures that the models remain accurate and reliable in predictions, guiding treatment decisions, and improving patient outcomes. Effective monitoring and adaptation strategies are essential for maintaining the relevance and effectiveness of machine learning applications in the management of predicting the disease.



Conclusion

In conclusion, this dataset presents an opportunity to create a machine learning model that can identify potential loan defaulters based on a variety of features. By effectively analyzing the dataset and addressing modeling challenges, participants can contribute to developing risk mitigation strategies for banks and financial institutions, ultimately supporting broader economic stability.

To setup Demo

Info.mlangles@cloudangles.com =

Visit: www.mlangles.ai