##### Description

**Background & Context**

The Thera bank recently saw a steep decline in the number of users of their credit card, credit cards are a good source of income for banks because of different kinds of fees charged by the banks like annual fees, balance transfer fees, and cash advance fees, late payment fees, foreign transaction fees, and others. Some fees are charged to every user irrespective of usage, while others are charged under specified circumstances.

Customers’ leaving credit card services would lead the bank to loss, so the bank wants to analyze the data of customers and identify the customers who will leave their credit card services and the reason for same – so that the bank could improve upon those areas

You as a Data Scientist at Thera Bank need to come up with a classification model that will help the bank improve its services so that customers do not renounce their credit cards

You need to identify the best possible model that will give the required performance

**Objective**

1. Explore and visualize the dataset.
2. Build a classification model to predict if the customer is going to churn or not
3. Optimize the model using appropriate techniques
4. Generate a set of insights and recommendations that will help the bank

**Data Dictionary:**

* CLIENTNUM: Client number. Unique identifier for the customer holding the account
* Attrition\_Flag: Internal event (customer activity) variable - if the account is closed then "Attrited Customer" else "Existing Customer"
* Customer\_Age: Age in Years
* Gender: Gender of the account holder
* Dependent\_count: Number of dependents
* Education\_Level:  Educational Qualification of the account holder - Graduate, High School, Unknown, Uneducated, College(refers to a college student), Post-Graduate, Doctorate.
* Marital\_Status: Marital Status of the account holder
* Income\_Category: Annual Income Category of the account holder
* Card\_Category: Type of Card
* Months\_on\_book: Period of relationship with the bank
* Total\_Relationship\_Count: Total no. of products held by the customer
* Months\_Inactive\_12\_mon: No. of months inactive in the last 12 months
* Contacts\_Count\_12\_mon: No. of Contacts between the customer and bank in the last 12 months
* Credit\_Limit: Credit Limit on the Credit Card
* Total\_Revolving\_Bal: The balance that carries over from one month to the next is the revolving balance
* Avg\_Open\_To\_Buy: Open to Buy refers to the amount left on the credit card to use (Average of last 12 months)
* Total\_Trans\_Amt: Total Transaction Amount (Last 12 months)
* Total\_Trans\_Ct: Total Transaction Count (Last 12 months)
* Total\_Ct\_Chng\_Q4\_Q1: Ratio of the total transaction count in 4th quarter and the total transaction count in 1st quarter
* Total\_Amt\_Chng\_Q4\_Q1: Ratio of the total transaction amount in 4th quarter and the total transaction amount in 1st quarter
* Avg\_Utilization\_Ratio: Represents how much of the available credit the customer spent

#### ****Submission Guidelines****

1. There are two ways to work on this project:

**i. Full-code way:**The full code way is to write the solution code from scratch and only submit a final Jupyter notebook with all the insights and observations.

**ii. Low-code way**. The low-code way is to use an existing solution notebook template to build the solution and then submit a business presentation with insights and recommendations.

The primary purpose of providing these two options is to allow learners to opt for the approach that aligns with their individual learning aspirations and outcomes. The below table elaborates on these two options.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Submission type | Who should choose | What is the same across the two | What is different across the two | Final submission file [IMP] | Submission Format |
| Full-code | Learners who aspire to be in hands-on coding roles in the future focussed on building solution codes from scratch | Perform exploratory data analysis to identify insights and recommendations for the problem | Focus on code writing: 10-20% grading on the quality of the final code submitted | Solution notebook from the full-code template submitted in .html format | .html |
| Low-code | Learners who aspire to be in managerial roles in the future-focussed on solution review, interpretation, recommendations, and communicating with business |  | Focus on business presentation: 10-20% grading on the quality of the final business presentation submitted | Business presentation in .pdf format with problem definition, insights, and recommendations | .pdf |

Please follow the below steps to complete the assessment. Kindly note that if you submit a presentation, ONLY the presentation will be evaluated. Please make sure that all the sections mentioned in the rubric have been covered in your submission.

**i. Full-code version**

* Download the full-code version of the learner notebook.
* Follow the instructions provided in the notebook to complete the project.
* Clearly write down insights and recommendations for the business problems in the comments.
* Submit only the solution notebook prepared from the learner notebook [format: .html]

**ii.** **Low-code version**

* Download the low-code version of the learner notebook.
* Follow the instructions provided in the notebook to complete the project.
* Prepare a business presentation with insights and recommendations to the business problem.
* Submit only the presentation [format: .pdf]

2. Any assignment found copied/plagiarized with other submissions will not be graded and awarded zero marks.

3. Please ensure timely submission as any submission post-deadline will not be accepted for evaluation.

4. Submission will not be evaluated if

* it is submitted post-deadline, or,
* more than 1 file is submitted.

#### ****Best Practices for Full-code submissions****

* The final notebook should be well-documented, with inline comments explaining the functionality of code and markdown cells containing comments on the observations and insights.
* The notebook should be run from start to finish in a sequential manner before submission.
* It is important to remove all warnings and errors before submission.
* The notebook should be submitted as an HTML file (.html) and NOT as a notebook file (.ipynb).
* Please refer to the FAQ page for common project-related queries.

#### ****Best Practices for Low-code submissions****

* The presentation should be made keeping in mind that the audience will be the Data Science lead of a company.
* The key points in the presentation should be the following:
  + Business Overview of the problem and solution approach
  + Key findings and insights which can drive business decisions
  + Business recommendations
  + Focus on explaining the key takeaways in an easy-to-understand manner.
  + The inclusion of the potential benefits of implementing the solution will give you the edge.
* Copying and pasting from the notebook is not a good idea, and it is better to avoid showing codes unless they are the focal point of your presentation.
* The presentation should be submitted as a PDF file (.pdf) and NOT as a .pptx file.
* Please refer to the FAQ page for common project-related queries.

 Scoring guide (Rubric) - Credit card Users Churn Prediction

| **Criteria** | **Points** |
| --- | --- |
| Exploratory Data Analysis and Insights - Problem definition, questions to be answered - Data background and contents - Univariate analysis - Bivariate analysis - Key meaningful observations on individual variables and the relationship between variables | 8 |
| Data pre-processing - Prepare the data for analysis - Feature Engineering - Missing value Treatment - Outlier Treatment Note: Please ensure no data leakage occurs among train-test and validation sets | 5 |
| Model Building - Original Data - Build 5 models (from decision trees, bagging and boosting methods) - Comment on the model performance \* You can choose NOT to build XGBoost if you are facing issues with the installation | 6 |
| Model building - Oversampled data - Oversample the train data - Build 5 models (from decision trees, bagging and boosting methods) - Comment on the model performance \* You can choose NOT to build XGBoost if you are facing issues with the installation | 6 |
| Model building - Undersampled data - Undersample the train data - Build 5 models (from decision trees, bagging and boosting methods) - Comment on the model performance \* You can choose NOT to build XGBoost if you are facing issues with the installation | 6 |
| Model Performance Improvement using Hyperparameter Tuning - Choose models that might perform better after tuning (tune at least 3 models out of 15 built in the previous steps) - Provide proper reasoning for tuning that model - Tune the best 3 models obtained above using randomized search and metric of interest - Check the performance of 3 tuned models | 13 |
| Model Performances and Final Model Selection - Compare the performance of tuned models - Choose the best model - Comment on the performance of the best model on the test set | 4 |
| Actionable Insights & Recommendations - Write down insights from the analysis conducted - Provide actionable business recommendations | 4 |
| Presentation / Notebook - Overall Quality - Structure and flow - Crispness - Visual appeal - Conclusion and Business Recommendations OR - Structure and flow - Well commented code - Conclusion and Business Recommendations | 8 |
| Points | 60 |

FAQ - Credit Card Users Churn Prediction

**1. How should one approach the Credit Card Users Churn Prediction project?**

* Before starting the project, please read the problem statement carefully and go through the criteria and descriptions mentioned in the rubric
* Then you should start with an exploratory analysis of the data.
* This understanding will help you identify the need for pre-processing the data.
* Once the data is ready, you can start with the steps that need to be followed as mentioned in the rubric
  + Build 6 models with original data
  + Build 6 models with oversampled data
  + Build 6 models with undersampled data
  + Choose 3 best models among 18 models built in the previous 3 steps
  + Tune 3 models
* It is important to close the analysis with key findings and recommendations to the business.

**2. I am trying to fit a model and getting this error:**

**ValueError: could not convert string to float: 'M'**

**How to resolve?**

Please check if the X\_train and X\_test consists of strings, and then create dummy variables using**pd.get\_dummies**

**3. I am getting this error while importing SMOTE even after successful installation of imblearn library:**

**ImportError: cannot import name 'delayed' from 'sklearn.utils.fixes' (C:\Users\anaconda3\lib\site-packages\sklearn\utils\fixes.py)**

**How to resolve?**

1. Run **!pip install delayed** in your Jupyter notebook.
2. Restart the kernel and try importing SMOTE again.

**4. I am getting this error while trying to tune random forest:**

**NotFittedError: All estimators failed to fit**

**How to resolve?**

The Numpy library might not be updated. You can update the Numpy library to the latest version using

**!pip install numpy==1.20.3**in your Jupyter notebook

OR

**pip install numpy==1.20.3**in Anaconda prompt

**5. Do we need to do anything with the income variable, mainly around the signs "K","$", "less"? Should we eliminate these and use a different range?**

The category names can be renamed but it is not necessary as it won't affect your model in any way.

**6. One column has "abc" values. Can I process by dropping these or replacing them with the most frequent values? Which one is recommended?**

Dropping the values is not suggested. You can treat them as missing and replace them using an appropriate method.

**7. I Did the capping method during EDA as shown in the supermarket campaign model notebook. But that was then capping the extreme outliers only. what about the remaining outliers. can I treat them after Splitting the data, before model building?**

For finding outliers you can use the IQR method. But if you want to remove them based on say IQR method (say 25 and 75 percentiles) then after splitting data your test set might have a different range than your training set for a particular feature for which you want to detect the outliers. In this case, after applying IQR your test set might not represent the training set very well and this can reduce the accuracy of the test set.

It's recommended to remove outliers before splitting the dataset. Treating outliers depends on the business problem we need to analyze whether they are outliers or the values that can be possible.

Outliers need to drop because they can make your model worse. According to the business problem if there are more outliers and cannot be dropped just use any transformations such as log or square root so that we can reduce their effect.

**8. Why I am getting a different number of columns after imputation and one-hot encoding?**

The extra column you are getting is due to a common mistake while using simple imputer.

Whenever we use simple imputer you should fit only once not multiple times.

It's imputing the value 'married' for the missing value of the education\_level column and while one-hot encoding it's creating a column for that.

The attached code is the correct way of doing

reqd\_col\_for\_impute = [*NAME OF COLUMNS WITH MISSING VALUES*]   
imputer = SimpleImputer(missing\_values=np.nan, strategy="most\_frequent")  
  
# Fit and transform the train data   
X\_train[reqd\_col\_for\_impute] = imputer.fit\_transform(X\_train[reqd\_col\_for\_impute])   
  
# Transform the validation data   
X\_val[reqd\_col\_for\_impute] = imputer.transform(X\_val[reqd\_col\_for\_impute])   
  
# Transform the test data   
X\_test[reqd\_col\_for\_impute] = imputer.transform(X\_test[reqd\_col\_for\_impute])

**9. How to deal with “ValueError: This solver needs samples of at least 2 classes in the data, but the data contains only 1 class”?**

The target variable is not encoded properly. The error shows that there is only 1 class. Ensure that you are encoding the target features properly.

**10. How to decide which approach should be taken to split the data into different sets and perform cross-validation?**

We can take the following two approaches two split the data:

Train/Test:

You can split the data into train and test. Train the model using the training set and report cross-validation on the train set using K-Fold cross-validation. Check the CV score to assess the performance and then use the test set to assess the performance only on the final model.

We should follow this approach when we do not have enough data to create three splits.

For eg, we have 500 data points, splitting the data into 60% train, 20% validation, and 20% test would result in having very limited data points in the train set and our model will not be able to identify the relevant patterns. Hence, train/test split might be an appropriate strategy in such cases.

Train/Validation/Test: We can split the data into train, test, and validation. Train the model using the train set, check the model performance on the validation set and tweak the hyperparameters by checking the performance on the validation set. Use test set to assess the performance only on the final model.

We should follow this approach when we have enough data to create three splits.

For eg, We have 10k data points, splitting the data into 60% train, 20% validation, and 20% test would result in having fair enough data points in the train set and our model might be able to identify the relevant patterns. Hence, we should make a train/test split in such cases.

Jan 14, 6:34 PM

Bagging tuned with oversampled data shows 0s in Recall, Precision, and F1 values. What could go wrong?

https://secure.gravatar.com/avatar/3075a90066701502c6b2952e6d508947?s=50&d=https%3A%2F%2Folympus.mygreatlearning.com%2Fimages%2Fmessages%2Favatar-50.png

**Evaluator**

Jan 14, 11:32 PM

The issue likely arises from how oversampling was applied. If oversampling (e.g., SMOTE) was performed before splitting the dataset, it can lead to data leakage, causing the model to overfit to duplicated synthetic samples and fail on unseen data. Another possibility is class imbalance persisting in the test set, affecting recall and precision. Ensure oversampling is applied only to the training set after splitting and verify the class distribution in the test set. Additionally, check if the bagging model's parameters (e.g., base estimator or random state) are configured correctly.

**Abhay Kumar Gunhalkar**

Jan 13, 10:29 PM

Wouldn’t the presence of missing values in a dataset impact the accuracy of Exploratory Data Analysis (EDA) if they are not imputed? Since to avoid data leakage, the missing values are imputed after the train validation test split ?

What is the right approach ?

https://secure.gravatar.com/avatar/3075a90066701502c6b2952e6d508947?s=50&d=https%3A%2F%2Folympus.mygreatlearning.com%2Fimages%2Fmessages%2Favatar-50.png

**Evaluator**

Jan 14, 11:32 PM

Missing values can impact EDA by skewing insights, especially in descriptive statistics or visualizations. However, imputing missing values before splitting risks data leakage. The right approach is to analyze the missingness during EDA to understand its pattern (e.g., missing completely at random or not). For EDA purposes, you can temporarily impute (e.g., mean, median) or exclude rows to visualize trends but avoid permanent imputations. After splitting into train-validation-test sets, impute missing values using only the training data's statistics to prevent leakage and maintain model integrity. Document all steps clearly for reproducibility.

https://secure.gravatar.com/avatar/29ab0c4d70cac3cf62235a51422f285a?s=50&d=https%3A%2F%2Folympus.mygreatlearning.com%2Fimages%2Fmessages%2Favatar-50.png

**Prasanth Pasala**

Jan 13, 1:10 AM

Should we remove sample parameter grid and sample tunning sections before submitting the html file?

https://secure.gravatar.com/avatar/3075a90066701502c6b2952e6d508947?s=50&d=https%3A%2F%2Folympus.mygreatlearning.com%2Fimages%2Fmessages%2Favatar-50.png

**Evaluator**

Jan 14, 11:32 PM

Not required you van keep it. Its up to you.

https://secure.gravatar.com/avatar/29ab0c4d70cac3cf62235a51422f285a?s=50&d=https%3A%2F%2Folympus.mygreatlearning.com%2Fimages%2Fmessages%2Favatar-50.png

**Prasanth Pasala**

Jan 12, 11:47 PM

###### In the below statement what does it mean by 3 models out of 15? There are 5 models each trained on 3 data sets (original, over sampled, under sampled). Does it mean just pick 3 out of DecisionTree, RandomForestClassifier, BaggingClassifier, AdaBoostClassifier, GradientBoosting, XGBoost or is it asking to pick 3 from the 15 combinations (5 models with 3 data sets) ?

###### From Rubric: ---------------------------------- Model Performance Improvement using Hyperparameter Tuning - Choose models that might perform better after tuning (tune at least 3 models out of 15 built in the previous steps) - Provide proper reasoning for tuning that model - Tune the best 3 models obtained above using randomized search and metric of interest - Check the performance of 3 tuned models \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

https://secure.gravatar.com/avatar/3075a90066701502c6b2952e6d508947?s=50&d=https%3A%2F%2Folympus.mygreatlearning.com%2Fimages%2Fmessages%2Favatar-50.png

**Evaluator**

Jan 14, 11:34 PM

 It is asking to pick 3 from the 15 combinations (5 models with 3 data sets).

https://secure.gravatar.com/avatar/e95ba6bc9bd15a25052a507b6959c68e?s=50&d=https%3A%2F%2Folympus.mygreatlearning.com%2Fimages%2Fmessages%2Favatar-50.png

**Ron Morgan**

Jan 12, 7:59 PM

for low code do we need to provide steps on how we handle outliers ?

###### Data pre-processing

- Prepare the data for analysis - Feature Engineering - Missing value Treatment - Outlier Treatment Note: Please ensure no data leakage occurs among train-test and validation sets

https://secure.gravatar.com/avatar/3075a90066701502c6b2952e6d508947?s=50&d=https%3A%2F%2Folympus.mygreatlearning.com%2Fimages%2Fmessages%2Favatar-50.png

**Evaluator**

Jan 12, 11:38 PM

No need of all steps, but technique and results should be mentioned.

https://secure.gravatar.com/avatar/a924090f8a6d4840c1ce0f93f9585d0c?s=50&d=https%3A%2F%2Folympus.mygreatlearning.com%2Fimages%2Fmessages%2Favatar-50.png

**Mathew Poon**

Jan 11, 12:33 PM

I am also getting error on Tuning XGBoost Model

https://secure.gravatar.com/avatar/595b0703a3110f010effc5e5ee7a0ac9?s=50&d=https%3A%2F%2Folympus.mygreatlearning.com%2Fimages%2Fmessages%2Favatar-50.png

**Liem Ba Nguyen**

Jan 06, 11:40 AM

I run XGboost, the system require to down grate to sklearn version 1.2.2 (current version 1.6). When I did that the result of other is slightly change.

https://secure.gravatar.com/avatar/3075a90066701502c6b2952e6d508947?s=50&d=https%3A%2F%2Folympus.mygreatlearning.com%2Fimages%2Fmessages%2Favatar-50.png

**Evaluator**

Jan 08, 6:37 AM

Downgrading to scikit-learn 1.2.2 may cause slight changes in results due to differences in underlying algorithms, default parameter values, or numerical precision between versions. Such changes are expected, as updates in scikit-learn often include optimizations and bug fixes that can subtly affect model behavior. To ensure consistency, document the exact library versions used during experimentation and consider re-tuning your model parameters in the downgraded environment to account for these variations.

https://secure.gravatar.com/avatar/f251950d2341b83ca2e3007146ed136a?s=50&d=https%3A%2F%2Folympus.mygreatlearning.com%2Fimages%2Fmessages%2Favatar-50.png

**Tahereh Hosseinimousa**

Jan 06, 8:24 AM

Seems XGBClassifier in SKLearning has API issue, not able to run it in current version.

https://secure.gravatar.com/avatar/3075a90066701502c6b2952e6d508947?s=50&d=https%3A%2F%2Folympus.mygreatlearning.com%2Fimages%2Fmessages%2Favatar-50.png

**Evaluator**

Jan 08, 6:38 AM

The API issue with XGBClassifier in the current scikit-learn version might stem from compatibility changes between scikit-learn and XGBoost. Such issues often arise when there are mismatches in expected parameter formats or updates in scikit-learn's model integration. To resolve this, consider checking the documentation for any breaking changes, updating XGBoost to a compatible version, or downgrading scikit-learn to an earlier version where the integration was stable. Ensure consistent library versions to avoid future compatibility issues.

https://secure.gravatar.com/avatar/e95ba6bc9bd15a25052a507b6959c68e?s=50&d=https%3A%2F%2Folympus.mygreatlearning.com%2Fimages%2Fmessages%2Favatar-50.png

**Ron Morgan**

Jan 05, 4:19 PM

getting this error when working on the xgboost section.   
# defining model  
Model = XGBClassifier(random\_state=1,eval\_metric='logloss')

#Parameter grid to pass in RandomSearchCV  
param\_grid={ 'n\_estimators': [50, 75, 100],  
    'scale\_pos\_weight': [1, 2],  
    'learning\_rate': [0.01, 0.1],  
    'gamma': [0, 1, 3],  
    'subsample': [0.7, 0.8, 0.9]  
           }  
from sklearn import metrics

# Type of scoring used to compare parameter combinations  
scorer = metrics.make\_scorer(metrics.recall\_score)

#Calling RandomizedSearchCV  
randomized\_cv = RandomizedSearchCV(estimator=Model, param\_distributions=param\_grid, n\_iter=50, n\_jobs = -1, scoring=scorer, cv=5, random\_state=1)  
#randomized\_cv = RandomizedSearchCV(estimator=Model, param\_distributions=param\_grid, n\_iter=5, n\_jobs=-1, scoring=scorer, cv=5, random\_state=1)

#Fitting parameters in RandomizedSearchCV  
randomized\_cv.fit(X\_train, y\_train) ## Complete the code to fit the model on original data

print("Best parameters are {} with CV score={}:" .format(randomized\_cv.best\_params\_,randomized\_cv.best\_score\_))

#print("Best parameters are {} with CV score={}:" .format(randomized\_cv.best\_params\_,randomized\_cv.best\_score\_))

---------------------------------------------------------------------------

ImportError Traceback (most recent call last)

<timed exec> in <module>

[/usr/local/lib/python3.10/dist-packages/xgboost/\_\_init\_\_.py](https://localhost:8080/) in <module>

5

6 from . import tracker # noqa

----> 7 from . import collective, dask, rabit

8 from .core import (

9 Booster,

[/usr/local/lib/python3.10/dist-packages/xgboost/dask/\_\_init\_\_.py](https://localhost:8080/) in <module>

64 from xgboost.\_typing import \_T, FeatureNames, FeatureTypes, IterationRange

65 from xgboost.callback import TrainingCallback

---> 66 from xgboost.compat import DataFrame, LazyLoader, concat, lazy\_isinstance

67 from xgboost.core import (

68 Booster,

ImportError: cannot import name 'concat' from 'xgboost.compat' (/usr/local/lib/python3.10/dist-packages/xgboost/compat.py)

https://secure.gravatar.com/avatar/3075a90066701502c6b2952e6d508947?s=50&d=https%3A%2F%2Folympus.mygreatlearning.com%2Fimages%2Fmessages%2Favatar-50.png

**Evaluator**

Jan 05, 11:46 PM

This error occurs due to an incompatibility between the installed version of xgboost and its internal dependencies. Specifically, the error indicates that the concat function is missing from the xgboost.compat module. This issue often arises with certain xgboost releases.

### Steps to Fix:

1. **Verify the XGBoost version**:
   * Run:
   * import xgboost
   * print(xgboost.\_\_version\_\_)
2. **Upgrade or Downgrade XGBoost**:
   * Upgrade XGBoost to a more recent and stable version where this issue is fixed:
   * pip install --upgrade xgboost
   * Alternatively, downgrade to a compatible version:
   * pip install xgboost==1.7.6
3. **Check Dependencies**: Ensure all other dependencies (like scikit-learn) are compatible with the XGBoost version you're using. For example:
   * For XGBoost ≥ 1.7.0, use scikit-learn ≥ 1.1.0.
4. **Clean Environment**: If upgrading/downgrading doesn't resolve the issue, create a new Python environment:
5. python -m venv myenv
6. source myenv/bin/activate # For Linux/macOS
7. myenv\Scripts\activate # For Windows
8. pip install xgboost scikit-learn
9. **Retry Your Code**: After fixing the versions, rerun your code to ensure the issue is resolved.

Let me know if further assistance is needed!

**Pablo Sandoval**

Jan 05, 3:01 PM

Hello - I'm getting an error on the "Tuning XGBoost Model with Original data" section when I run the code. See attached images. Please advise.

Jan 05, 3:01 PM

Here is the code screenshot

Screenshot 2025-01-05 145916 Code.png

https://secure.gravatar.com/avatar/3075a90066701502c6b2952e6d508947?s=50&d=https%3A%2F%2Folympus.mygreatlearning.com%2Fimages%2Fmessages%2Favatar-50.png

**Evaluator**

Jan 05, 11:48 PM

The error occurs due to incompatibility between XGBClassifier and the scikit-learn version, as RandomizedSearchCV relies on certain internal methods (e.g., \_\_sklearn\_tags\_\_). To fix it:

### Steps:

1. **Check Compatibility**:
   * Ensure xgboost and scikit-learn versions are compatible:
     + For xgboost ≥ 1.7.0, use scikit-learn ≥ 1.1.0.
     + For xgboost ≤ 1.6.2, use scikit-learn ≤ 1.0.2.
2. **Update or Downgrade Libraries**:
   * Run:
   * pip install xgboost==1.7.6 scikit-learn==1.2.2
3. **Modify the Code**:
   * Replace metrics.recall\_score with average='binary' if using multi-class data.
4. **Validate Installation**:
   * Test imports:
   * from xgboost import XGBClassifier
   * from sklearn.model\_selection import RandomizedSearchCV
5. **Retry the Code**: Ensure you are fitting on compatible data (X\_train, y\_train).

By following these steps, the issue should be resolved. Let me know if you need further assistance!

**Pablo Sandoval**

Jan 02, 7:33 PM

Hello - I'm working on the low code and I noticed that on the univariate analysis, the instructions are showing ## Complete the code to create histogram\_boxplot for 'New\_Price' for several of the histogram\_boxplot. Please advice.

https://secure.gravatar.com/avatar/3075a90066701502c6b2952e6d508947?s=50&d=https%3A%2F%2Folympus.mygreatlearning.com%2Fimages%2Fmessages%2Favatar-50.png

**Evaluator**

Jan 03, 12:01 AM

Hello, in low code version, it is important that you show crucial plots (like boxplot, histograms etc). No need to show the code.

(1)

https://secure.gravatar.com/avatar/9ccf27034513c7065e1ae56535baa814?s=50&d=https%3A%2F%2Folympus.mygreatlearning.com%2Fimages%2Fmessages%2Favatar-50.png

**Mini Francis**

Jan 17, 1:21 PM

Hi, I have query about similar thing. In full code version there is 6 questions to be answered as part of  EDA analysis.  Do we need to do each graph with observation as well?

https://secure.gravatar.com/avatar/327ecfcb4d554375d2bb3651f21a166f?s=50&d=https%3A%2F%2Folympus.mygreatlearning.com%2Fimages%2Fmessages%2Favatar-50.png

**John P. Bossert**

Dec 26, 2024, 1:31 PM

How are you addressing the missing values for (1) Education\_Level and (2) Marital\_Status ?

https://secure.gravatar.com/avatar/3075a90066701502c6b2952e6d508947?s=50&d=https%3A%2F%2Folympus.mygreatlearning.com%2Fimages%2Fmessages%2Favatar-50.png

**Evaluator**

Dec 28, 2024, 12:59 AM

To address missing values for the variables **"Education\_Level"** and **"Marital\_Status"**, the following strategies can be applied:

1. **Analyze Missing Data**:
   * Use .isnull().sum() to identify the number of missing values.
   * Check if the missingness is random or correlated with other variables using statistical tests or visualizations (e.g., heatmaps).
2. **Imputation Techniques**:
   * For **"Education\_Level"** (categorical):
     + Impute missing values with the **most frequent category** (mode) if the variable is ordinal.
     + Alternatively, impute based on patterns in other variables using techniques like **K-Nearest Neighbors (KNN)** or **decision trees**.
   * For **"Marital\_Status"** (categorical):
     + Impute with the **mode** or group-wise mode (e.g., based on age or income groups).
     + Use similar advanced techniques if relationships with other variables exist.
3. **Advanced Imputation**:
   * If a significant portion of data is missing, consider creating a **"missing" category** to retain missingness information.
   * For predictive models, use algorithms that handle missing values directly (e.g., XGBoost).
4. **Validate Impact**:
   * Compare results before and after imputation to ensure the approach doesn't distort the dataset.

**Sai Donepudi**

Dec 26, 2024, 9:16 AM

When we do Hyperparameter tuning, do we use GridSearchCV or RandomizedSearchCV?

https://secure.gravatar.com/avatar/3075a90066701502c6b2952e6d508947?s=50&d=https%3A%2F%2Folympus.mygreatlearning.com%2Fimages%2Fmessages%2Favatar-50.png

**Evaluator**

Dec 28, 2024, 12:58 AM

Use **GridSearchCV** when you want to exhaustively search all possible combinations of hyperparameters, typically for smaller parameter spaces. Use **RandomizedSearchCV** for larger parameter spaces or when computational resources are limited, as it randomly samples a fixed number of combinations. RandomizedSearchCV is faster and can still find near-optimal parameters. Choose based on the size of the parameter grid and available resources.

(1)

**Merlin Balamurugan**

- EDA is an important part of any project involving data.

- It is important to investigate and understand the data better before building a model with it.

- A few questions have been mentioned below which will help you approach the analysis in the right manner and generate insights from the data.

- A thorough analysis of the data, in addition to the questions mentioned below, should be done.

**\*\*Questions\*\***:

1. How is the total transaction amount distributed?

2. What is the distribution of the level of education of customers?

3. What is the distribution of the level of income of customers?

4. How does the change in transaction amount between Q4 and Q1 (`total\_ct\_change\_Q4\_Q1`) vary by the customer's account status (`Attrition\_Flag`)?

5. How does the number of months a customer was inactive in the last 12 months (`Months\_Inactive\_12\_mon`) vary by the customer's account status (`Attrition\_Flag`)?

6. What are the attributes that have a strong correlation with each other?