

CREDIT SCORE CLASSIFICATION GUIDELINE

JUN ZE HE

Introduction: Credit Score Classification Model Credit Score is an important factor that is how banks determine how much credit balance you can borrow from your credit account. To save manual efforts to identify whether a credit account has a good credit score, we want to develop a classification model, that can classify credit scores from peoples' credit accounts.

[Credit Score Dataset](#)

Week3: What is a classification model

Goal

- (1) Get familiar with Google colab
- (2) Know how to upload files and connect to your Google Drive on Google colab
- (3) Take a look at my data science functions repository and know how to use functions
- (4) recognize what the predicted variables and the response variable are in the dataset
- (5) Get to know the basics of machine learning

Binary classification has two classes, but multi-class classification has 2 or more than two classes

Guide to upload files and use my data science functions

- (1) download the dataset from Kaggle and upload the dataset to your colab notebook folder in your Google Drive
- (2) click the folder button on the left side
- (3) Under the name "Files", there are four tabs, and press the third button to mount your Google Drive
- (4) download my data science functions from my repository and upload it to the content directory on Google colab

■

Resources

[What is machine learning, and what is a classification model?](#)

[My Data Science Functions Repository](#)

[The Project Repository](#)

■

Week4: Data Cleaning

Goal

The goal of data cleaning is to clean the dataset without NULL and strange values

- (1) First step: check if there is any value in the correct form, for example, price is an object, any variable has an incorrect value
- (2) Second step: identify how many null values are in the dataset, then try to correct them with median or mean in numerical variables and most frequent values in categorical variables

■

Resources

[Pandas Tutorials](#)

[numpy Tutorials](#)

[Handling missing values in categorical variables](#)

[Handling missing values in general](#)

■

Week5: statistical analysis / EDA**Goal**

EDA

- (1) read the EDA for classification models article and follow the instructions to analyze data
- (2) find the relationship between each of the numerical variables and categorize them by the response variable (scatter plot)
- (3) encodes the categorical variables by ordinal numbers to find the relationship between each of the categorical variables and categorizes them by the response variable (scatterplot)
- (4) find the correlation between each variable to make sure there is no multicollinearity problem and plot the correlation value with a heatmap

statistical analysis

- (1) read Logistic Regression Using statsModels article and implement codes from Multinomial Logistic Regression article
- (2) know what is the statsmodel package and how to use it
- (3) know how to check the significance of the model and variables
- (4) check marginal effects
- (5) check variance inflation factor

■

Resources[EDA for classification models](#)[Multinomial Logistic Regression](#)[Logistic Regression Using StatsModels](#)[variance inflation factor](#)

■

Week6: Feature Selection / Feature Engineering**Goal**

■

Resources [A complete feature selection guide](#)

■

Week7: Data Preprocessing / Models / Evaluations**Goal**

- (1) oversampling, undersampling, and manually reducing the majority group in the predicted variable
- (2) encode categorical variables
- (3) train-test split
- (4) standardization (check if we should oversample the data before the standardization or after the standardization by statsmodel)
- (5) models: RandomForestClassifier, XGBClassifier, LogisticRegression, SupportVectorMachine (know what are the advantages of these models and what are the drawbacks)
- (6) evaluate the models by validation scores, F1-score, Specificity, and recall-score

■

Resources[RandomForestClassifier](#)[XGBClassifier](#)[Logistic Regression](#)[Support Vector Machine](#)[Evaluation Metrics \(F1-Score, Recall, Precision\)](#)[Confusion Matrix](#)

■

Week8: Conclusions/Making slides

Goal

