# Model Card Project Deliverable (40 pts.):

In your new Github repository, use [GitHub Markdown](https://guides.github.com/features/mastering-markdown/) to transform your README.md file into a model card. (Each group only needs one repository.) Ensure your model card follows these bullet points:

# Basic information (6 pts.):

* + *Group member name*: Pawan Kumar , email:[pawanmarine81@gwu.edu](mailto:pawanmarine81@gwu.edu)
  + *Date*: 26-August-2022
  + Model version: Python version: 3.7.13, sklearn version: 1.0.2
  + *License*: MIT
  + *Model implementation code*: <https://github.com/jphall663/GWU_DNSC_6301_project/blob/main/DNSC_6301_Example_Project.ipynb>
  + *i) Intended Use*: By learning this model, it can use as tool of improve the chances of improving the credit line by figure out the probability of defaulter.
  + ii) *Intended users*: Mostly by the Graduate student, especially the data and Business analysts’ students
  + iii) *Out-of-scope uses*: Used in context other than educational purpose.

# Training data (8 pts.):

* *Source of training data*: Data is available in GWU course Analytical Edge and Data ethics \_DNSC\_6301\_11
* *How training data was divided into training and validation data*: Using Training, test and split, in which 50% training, 25% test, and 25% Validation
* *Number of rows in training and validation data*: 15000 rows, 20 columns in training data, whereas in validation there are 7500 rows, 20 Columns.
* Data dictionary: for each column in the training dataset include:

| ***Name*** | ***Modeling Role*** | ***Measurement Level*** | ***Description*** |
| --- | --- | --- | --- |
| **ID** | ID | int | unique row identifier |
| **LIMIT\_BAL** | input | float | amount of previously awarded credit |
| **SEX** | demographic information | int | 1 = male; 2 = female |
| **RACE** | demographic information | int | 1 = Hispanic; 2 = black; 3 = white; 4 = Asian |
| **EDUCATION** | demographic information | int | 1 = graduate school; 2 = university; 3 = high school; 4 = others |
| **MARRIAGE** | demographic information | int | 1 = married; 2 = single; 3 = others |
| **AGE** | demographic information | int | age in years |
| **PAY\_0, PAY\_2 - PAY\_6** | inputs | int | history of past payment; PAY\_0 = the repayment status in September 2005; PAY\_2 = the repayment status in August 2005; ...; PAY\_6 = the repayment status in April 2005. The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; ...; 8 = payment delay for eight months; 9 = payment delay for nine months and above |
| **BILL\_AMT1 - BILL\_AMT6** | inputs | float | amount of bill statement; BILL\_AMNT1 = amount of bill statement in September 2005; BILL\_AMT2 = amount of bill statement in August 2005; ...; BILL\_AMT6 = amount of bill statement in April, 2005 |
| **PAY\_AMT1 - PAY\_AMT6** | inputs | float | amount of previous payment; PAY\_AMT1 = amount paid in September 2005; PAY\_AMT2 = amount paid in August 2005; ...; PAY\_AMT6 = amount paid in April, 2005 |
| **DELINQ\_NEXT** | target | int | whether a customer's next payment is delinquent (late), 1 = late; 0 = on-time |

# Test data (5 pts.):

* *Source of test data*: Data is available in GWU course Analytical Edge and Data ethics \_DNSC\_6301\_11 (By splitting data in to training and testing data)
* *Number of rows in test data*: 7500 rows, 20 columns
* *State any differences in columns between training and test data:* There is no difference

# Model details (8 pts.):

* *Columns used as inputs in the final model:* 'LIMIT\_BAL', 'PAY\_0', 'PAY\_2', 'PAY\_3', 'PAY\_4', 'PAY\_5', 'PAY\_6', 'BILL\_AMT1', 'BILL\_AMT2', 'BILL\_AMT3', 'BILL\_AMT4', 'BILL\_AMT5', 'BILL\_AMT6', 'PAY\_AMT1', 'PAY\_AMT2', 'PAY\_AMT3', 'PAY\_AMT4', 'PAY\_AMT5', 'PAY\_AMT6']
* *Column(s) used as target(s) in the final model*: ‘DELINQ\_NEXT'
* *Type of model*: Decision tree
* *Software used to implement the model:* Python, Pandas, and Sklearn
* *Version of the modeling software*: sklearn version: 1.0.2, Python version: 3.7.13

*Hyperparameters or other settings of your model:* DecisionTreeClassifier(ccp\_alpha=0.0, class\_weight=None, criterion='gini',

max\_depth=6, max\_features=None, max\_leaf\_nodes=None,

min\_impurity\_decrease=0.0, min\_impurity\_split=None,

min\_samples\_leaf=1, min\_samples\_split=2,

min\_weight\_fraction\_leaf=0.0, presort='deprecated',

random\_state=12345, splitter='best’) `

# Quantitative analysis (7 pts.):

* + *Metrics used to evaluate your final model (AUC and AIR)*: AUC: 0.7438, AIR from model are 0.82, 0.83 and 0.98 for Hispanic to white, black to white, and Asian to white respectively, and AIR values for recalculated model are 0.84, 0.87, and 0.99 for Hispanic to white, black to white , and Asian to white respectively.
  + *State the final values, neatly -- as bullets or a table, of the metrics for all data: training, validation, and test data:*

Table

Description automatically generated

* *Provide any plots related to your data or final model* -- be sure to label the plots!

Chart, line chart

Description automatically generated

# *Ethical considerations (6 pts.):*

* + ***Describe potential negative impacts of using your model****:*

*Math/Software Problems*: There is always a high risk of attaining inconsistent results if results are based on recent payment trend while the long-term trends of are overlooked. One of the software problems. In this model also the variable is carrying significant importance too.

*Real world risks*: Because of inconsistent result, it is likely that the customers’ credit limit may remain plateau, if not dipped, i.e., increase in credit limits of customer’s’ are almost unlikely. Here the AIR value of Hispanic-to-white is 0.82 which is quite reasonable.

***Describe potential uncertainties relating to the impacts of using your model:***

*Math/Software Problems*: A classic example of software problem is that It is quite possible because of one or few low metrics among several high scoring metrics may cause the a unjustified result, eventually causing low credit limit for customers.

*Real world risks*: If the output of model will be implemented by the users, then the customers will not be granted additional credit what they deserve. In bigger prospect it will not only hamper the an Individual but also will adversely impact the economy, a negative aftermath of bias model.