Scoring Model for Overall performance of Student

Arnab Jana

June 25, 2020

Introduction:

The purpose of this project is to design a model to calculate the overall score of a student based on his/her characteristic features. This model could then be used to rank students and categorize them into groups, which could further be used for study or selection.

Description of data:

In this project, I have used the following features for each student.

1. Academics
2. Sports
3. Fitness
4. Social
5. Extra-curricular activities
6. Sleeping habits
7. Time spent on social media
8. Time spent on outdoor activities
9. height

Each feature is a categorical variable, taking 5 distinct values (1-5). **The score of a student is a continuous variable** (mostly concentrated between 100 and 500). For the purpose of this experiment, I have used a training dataset consisting of 2000 students.

The first 5 rows of the dataset are as follows:-

academics,sports,fitness,social,extra-currics,sleeping,time-on-social-media,time-on-outdoor,height
5, 4, 3, 5, 5, 4, 1, 3, 1, 401.9367354144325
5, 4, 2, 4, 2, 5, 3, 1, 2, 342.23034612763684
3, 5, 1, 2, 5, 2, 1, 1, 4, 281.00220669805213
1, 2, 4, 4, 5, 1, 5, 3, 1, 203.99457657370246
3, 5, 5, 4, 1, 1, 3, 1, 4, 321.87808309204877

Generation of data:

I have assigned certain weights to every category of each of the 9 features (putting highest emphasis on academics and other features following obvious trends). The score, however, does not depend on outdoor activities and height. While creating an entry for a student, **I randomly assign the values of each feature (1-5) and the overall score is the summation of weights corresponding to each category perturbed by a standard normal distribution.** I have intentionally created some outliers where the perturbation is from a gaussian with variance 10 and mean which is drawn from a uniform (-50,+50) distribution.

Feature Selection:

I have experimented a few techniques for feature selection. The motivation behind this are [1]:

1. The response variable may be independent of a lot of features
2. Some features may be highly correlated among each other. Next, I describe the results of each of them.
3. Dimensionality reduction: Reduces model complexity and training time.

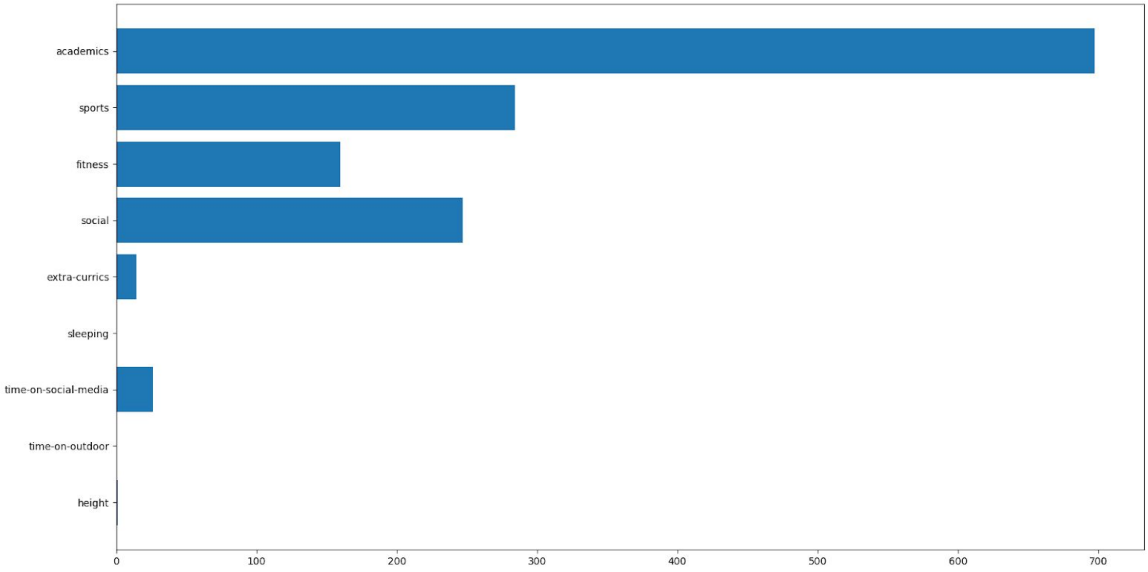
The major intuition for most of the algorithms is that if for different values of a particular feature, the values of the response variable are cluttered together, then the variable is almost independent of the feature. So, the larger the variance of the response variable w.r.t a feature, the larger is the importance of the feature.

Now, I will briefly describe some of the feature selection methods and their outcomes. Note that in our case, the predictor variables are categorical and the response variable is continuous.

1. ANOVA (Analysis Of Variance)

   Multifactor ANOVA [2] tests the relationship between categorical predictors vs continuous response. If there is equal variance between groups of categorical feature w.r.t continuous response, then we can conclude that this feature has no impact on response and hence should not be used for model training.

   In our experiment, the result of performance of ANOVA on the 9 features gives the F-scores. Higher F-scores indicates dependence of the feature on the response variable.



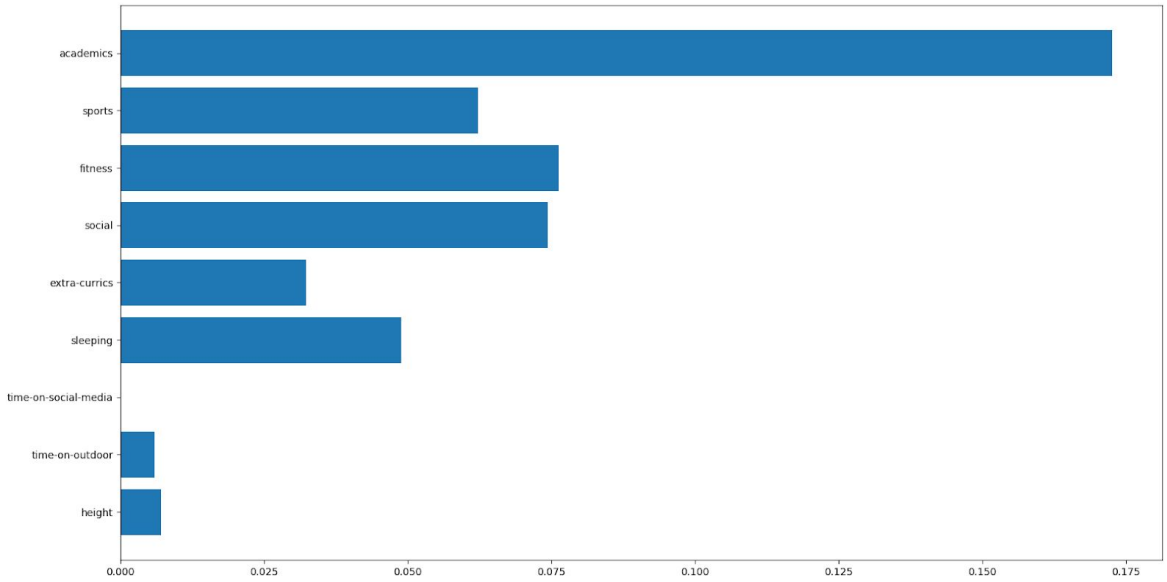
   From the bar-graph, we can conclude that sleeping, outdoor activities and height have no effect on the overall score of a student.

2. Mutual Information (MI)

   Mutual information [3] of two random variables is a measure of the mutual dependence between the two variables. More specifically, it quantifies the "amount of information" obtained about one random variable through observing the other random variable. **The concept of mutual information is intricately linked to that of entropy of a random variable** (more the entropy, more the amount of information held in a random variable).

Not limited to real-valued random variables and linear dependence like the correlation-coefficient, MI is more general and determines how different the joint distribution of the pair (X,Y) is to the product of the marginal distributions of X and Y.

In our experiment, we calculate the MI between each feature and the response variable.



From the bar-graph, we can conclude that social-media, outdoor activities and height have negligible impact on the overall score of the student.

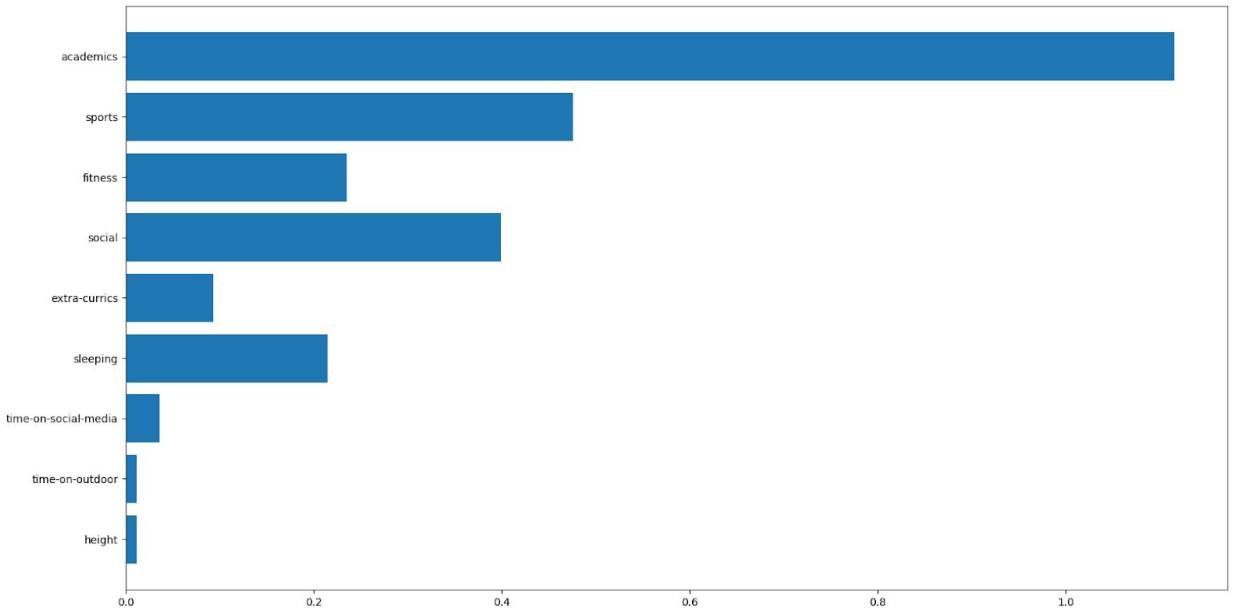
3. Information Value (IV)

Information Value [4] is an effective feature selection technique in a predictive model. It is calculated according to the following formula:-

$$\text{IV:} \quad \sum_{i=1}^{n} (\text{Distr Good}_i - \text{Distr Bad}_i) * ln\left(\frac{\text{Distr Good}_i}{\text{Distr Bad}_i}\right)$$

Here, Dist Good_i and Dist Bad_i are the fraction of good samples out of all the good samples and the fraction of bad samples out of all the bad samples respectively, participating in the i_th value of the categorical feature. Note that IV value is always positive and it is more when the distribution of good and bad samples inside a category differs significantly.
In our experiment, I have divided the continuous response variable into 2 categories, splitting approximately at the midpoint. The higher scores belong to

“good category” and vice-versa. The IV for each of the 9 features is plotted below.



From the bar-graph, we conclude that outdoor activities and height have negligible impact on the overall score of the student.

Result:

All the 3 feature selection methods give highest priority to academics and high priorities to sports and social. This is in accordance with the distribution of our generated dataset. The difference lies in the identification of the least important features.
We can see the Information Value (IV) method is closest to our underlying distribution, because our generated score was indeed independent of outdoor activities and height. **Even though MI and ANOVA also give low importance to these features, the best match for IV can be attributed to the fact that weights (in training data) are drawn from a distribution (not static) and there are outliers too, hence binning into classes (here 2) performs some kind of regularization**. Intuitively, it doesn't matter how much lower the score of a tall student is to a shorter student, as long as both have a low score (“low” based on threshold).

Mathematical model and scoring mechanisms.

In this experiment, we follow a scorecard approach, where the value in each categorical feature of a student results in a score and the overall score of a student is the sum of all individual scores.

Categorical feature	**Category**	**Score**
Academics	1	-30
Academics	2	+40
Sports	1	+20
Sports	2	-10

A sample scorecard where there are 2 categorical features, each having 2 categories (1 and 2).

Our job is to learn the score for each category, which most accurately predicts the overall score of the student.
Next, I describe a few approaches used in this experiment to learn this scorecard.

1. Weight of Evidence (WoE) [5]
   Here, we don't learn the score for each category of each feature. Instead, we calculate the score of each category from the training data (called Weight of Evidence) and then try to find the appropriate scaling factor for each category and the overall bias.

   $$\text{WoE:} \qquad \left[\ln\left(\frac{\text{Distr Good}}{\text{Distr Bad}}\right) \right] \times 100.$$

   Refer to Information value in the section Feature Selection to know about Distr Good and Distr Bad.

After calculating WoE for every category of each feature, the problem that remains is to find the appropriate scales and biases. This is equivalent to solving a Linear Regression Problem with WoE in place of feature categories. To perform regularization, I have used Ridge regression with alpha=0.1

Observations:
No. of parameters= no. of features+1= 8
Mean squared error on test-data: 4.81
Accuracy after discretized binning:

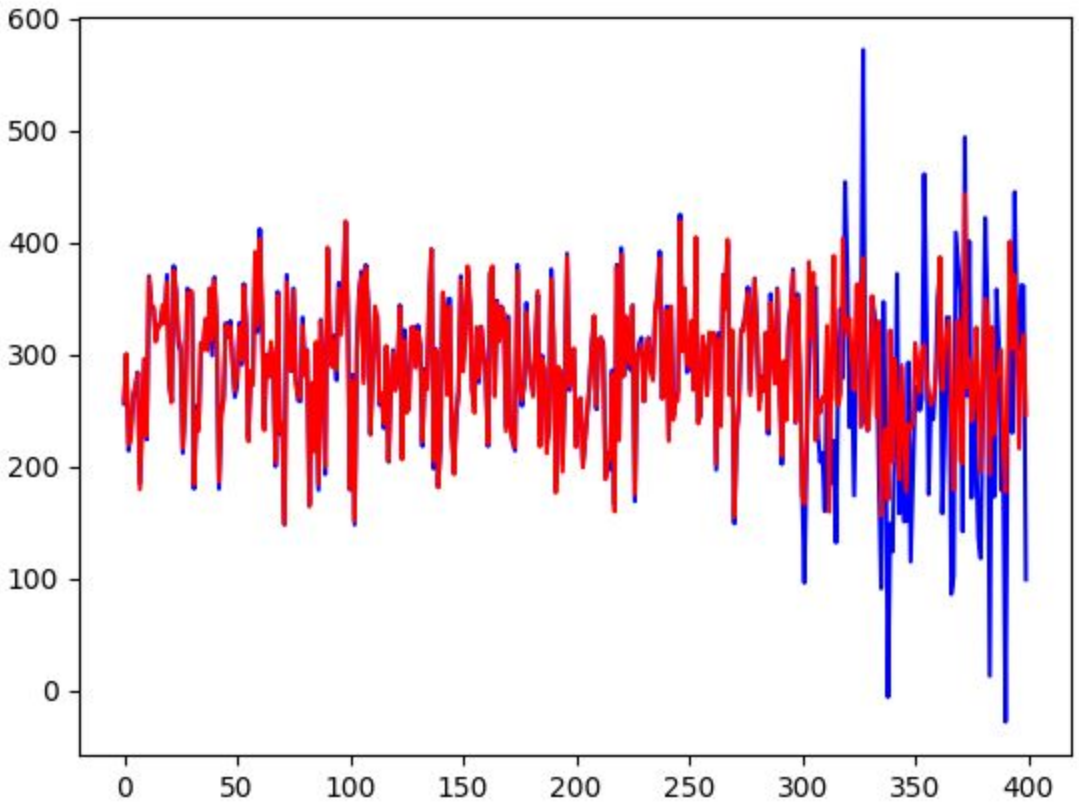
No. of bins	Accuracy (in %)
2	99.75
3	89.0
4	90.25
5	85.25
6	83.0
7	80.25
8	77.5
9	75

2. Regularized Linear Regression

In this model, we attempt to learn the score for each category for every feature in an end-to-end fashion. We first turn each categorical feature into its one-hot encoding. This flattened list of one-hot encodings forms the feature set. We the perform Ridge regression with alpha=0.1

Observations:
No. of parameters= (no.of features*no. of categories) +1= 36
Mean squared error on test-data: 4.62
Accuracy after discretized binning:

No. of bins	Accuracy (in %)
2	99.75
3	92.0
4	90.75
5	88.25
6	85.75
7	83.5
8	82.5
9	80.0

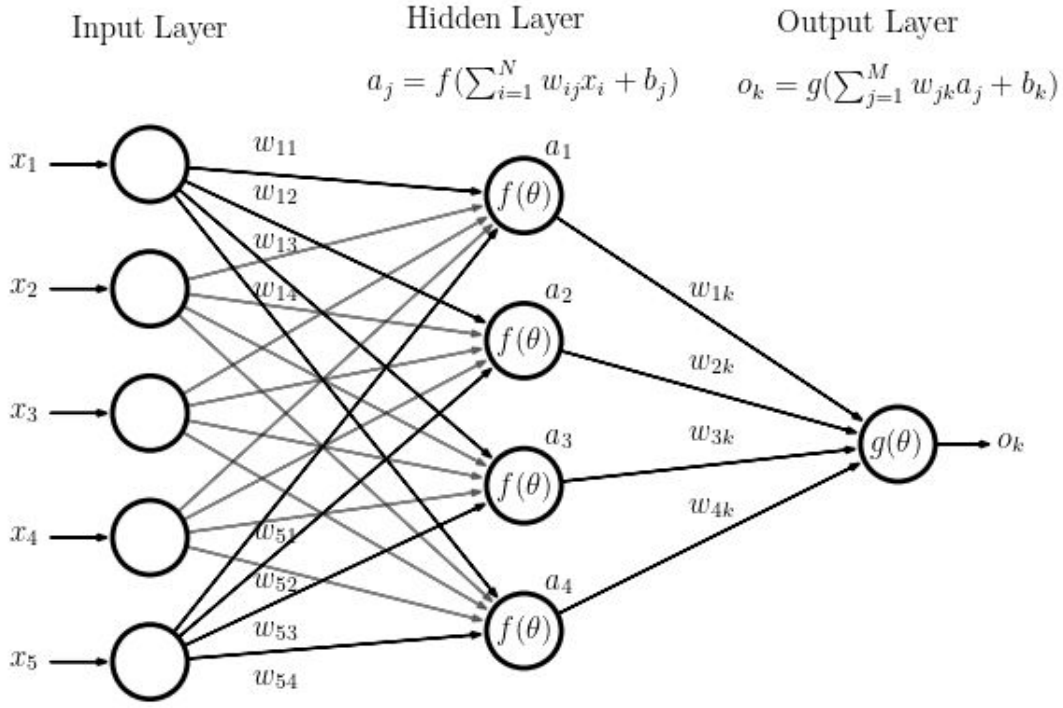


The blue and red lines indicate the test labels and predicted labels respectively. The samples in the rightmost ¾-th fraction are outliers and our model performs relatively well without overfitting.

3. Neural Network

The previous case assumes that there is a linear relationship between individual scores and overall score. To make it more generalized and adaptable to arbitrary nonlinear functions, we introduce a neural network to learn the weights corresponding to one-hot encoded input feature set (as in Regularized Linear Regression case).

The architecture consists of an input layer of size (no. of features*no. Of categories=35), a dense hidden layer of size 15 and a single output unit. Activation function: Rectified Linear Unit (ReLU). The model is trained with Mean Squared Error loss-function and mini-batch Adam optimizer.



Diagrammatic representation of the approximate structure of Neural Network used

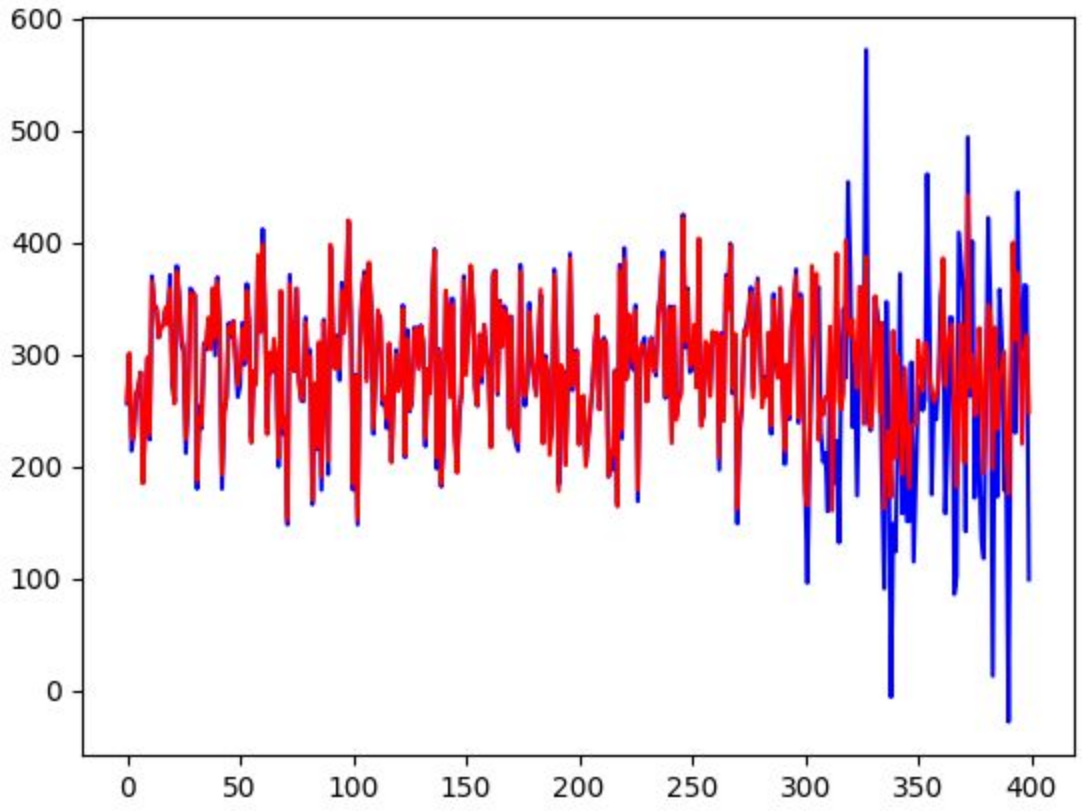
Observations:

No. of parameters= 35*15+15*1= 540

Mean squared error on test-data: 4.71

Accuracy after discretized binning:

No. of bins	Accuracy (in %)
2	99.75
3	92.0
4	90.5
5	88.5
6	85.25
7	83.5
8	80.75
9	80.0



The blue and red lines indicate the test labels and predicted labels respectively. The samples in the rightmost ¾-th fraction are outliers but our model tries to overfit to the outliers.

Result:

We can see that end-to-end models dominate over the Weight of Evidence method. Both Linear Regression and Neural Networks perform well on our data. **Linear regression marginally dominates NN. This can be attributed to the fact that our underlying data was mostly linear with some gaussian perturbations and NN might have a tendency to overfit due to non-linear activations.** However, depending on the underlying data-distribution, NN is expected to perform better in a large no. of scenarios.

Alternate scoring methods in literature:

Another scoring approach could be to use **logistic regression** or **random forest** (ensemble of decision trees) instead of a scorecard approach. These are extensively used in the FICO credit scoring system.

However, these are **more suited in a binary classification problem like credit risk analysis**, where the target variable (during training) is whether a customer has defaulted or not. The parameters learnt are then appropriately scaled into a credit score range, during inference on test data.

This method does not work well for regression or multi-class classification scenarios. An alternate experiment has been tried by discretizing the continuous target variable into 2 bins (good and bad) and then applying logistic regression, however, this method involves significant oversimplification and the performance is significantly poor.

Conclusion:

We can use Information Value (IV) or ANOVA for feature selection by removing features with negligible contribution. Then the feature set can be squashed to a one-hot encoding and scores for each category can be trained using Ridge Regression or artificial Neural Network.

The code has been written in Python3. The training data used can be found in data.csv. Alternatively the script to generate own training data is also available. The scripts to run feature-selection and model-training are attached and can be used to easily reproduce the results of the experiments above.

Links and References:

1. https://towardsdatascience.com/the-5-feature-selection-algorithms-every-data-scientist-need-to-know-3a6b566efd2
   Motivation behind Feature selection
2. https://towardsdatascience.com/anova-for-feature-selection-in-machine-learning-d9305e228476
   Anova and its variants
3. https://en.wikipedia.org/wiki/Mutual_information
   Mutual Information
4. https://www.listendata.com/2015/03/weight-of-evidence-woe-and-information.htm
   Information Value
5. https://medium.com/henry-jia/how-to-score-your-credit-1c08dd73e2ed
   Weight of Evidence and Information Value
6. http://www.m-hikari.com/ams/ams-2013/ams-9-12-2013/nurlybayevaAMS9-12-2013.pdf
   Paper on Algorithmic scoring models (Discriminant analysis, linear regression, logistic regression, Decision tree, NN)
7. https://www.berkeleybridge.com/blog/what-are-the-scoring-models-and-how-do-they-come-about/
   Introduction about scoring models.
8. https://medium.com/henry-jia/how-to-score-your-credit-1c08dd73e2ed
   Credit scoring- scorecard approach.