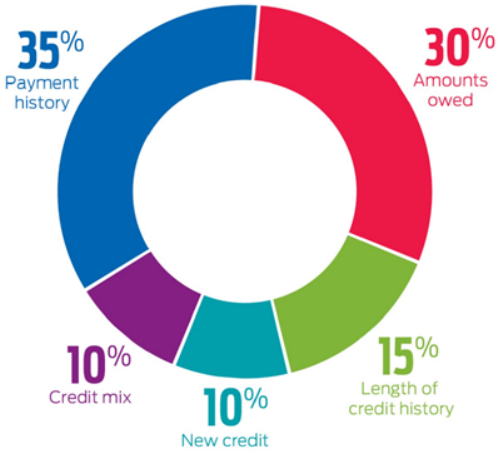
STAT3612 Statistical Machine Learning: Test 1

October 11, 2019 (50min)

Name: _____ UID: _____ Score: _____

Q1. (Understanding ScoreCard)
Shown below is a partial example of FICO ScoreCard in assessing personal credit score ranging from 350 to 800. On the left is the breakdown points per category and characteristics. On the right is the weight assignment for each category. Answer the following questions based on what you have learned from STAT3612.

Category	Characteristics	Attributes	Points
Payment History	Number of months since the most recent serious delinquency	No serious delinquency	75
		0 – 5	10
		6 – 11	15
		12 – 23	25
		24+	55
Outstanding Debt	Overall utilization on revolving trades	No revolving trades	30
		Under 6%	65
		7 – 19%	50
		20 – 49%	45
		50 – 89%	25
		90% or more	15
Credit History Length	Number of months in file	Below 12	12
		12 – 23	35
		24 – 47	60
		48 or more	75
Pursuit of New Credit	Number of inquiries in the last 6 months	0	70
		1	60
		2	45
		3	25
		4+	20
Credit Mix	Number of bankcard trade lines	0	15
		1	25
		2	55
		3	60
		4+	50



Source: FICO

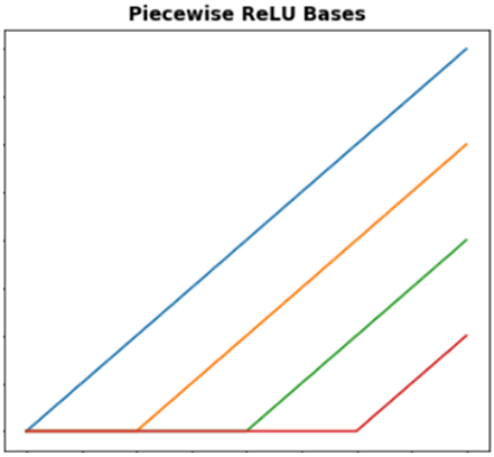
a) (2 marks) It is known the logistic regression model is used for predicting a *Good* or *Bad* credit. Assume the intercept is zero. Write down the model form in terms of category variables, together with their estimated coefficients.

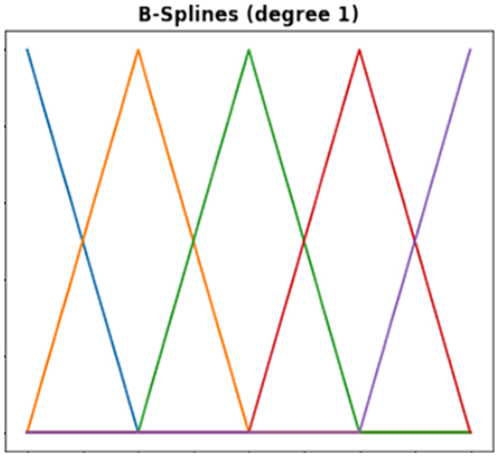
b) (2 marks) It is known the binning technique is used for capturing the nonlinear relationships between category characteristics and binary responses. Based on the ScoreCard charts, draw the binning-resulted step functions for *Credit History Length* and *Credit Mix*.

c) (2 marks) It is known the breakdown score points are linearly transformed from the predicted log-odds. Interpret the points (75, 10, 15, 25, 55) associated with the payment history.

d) (2 marks) Based on the ScoreCard charts, make your recommendations in terms of each category variable for improving the final credit score.

Q2. (ReLU vs. Linear B-Splines)
In the chapter of feature engineering, both ReLU and Linear B-Splines (LBS) are introduced as useful bases for piecewise linear regression. Given the same knots, their look differently as below. Suppose we have the sufficient data for each piece of intervals. Answer the following questions.





a) (2 marks) Justify whether the predicted piecewise curve $\hat{f}_{\mathrm{ReLU}}(x)$ is continuous over the entire x-domain. Justify whether so is $\hat{f}_{\mathrm{LBS}}(x)$.

b) (2 marks) Justify whether $\hat{f}_{\mathrm{ReLU}}(x)$ is identical to $\hat{f}_{\mathrm{LBS}}(x)$ for any input x.

c) (2 marks) What are the advantages of using the LBS bases compared to the use of ReLU? (Hint: multiple collinearity.)

Q3. (Smoothing Spline as a generalized ridge estimator)
The smoothing spline is a popular nonparametric smoother subject to ℓ_2-regularization, formulated as

$$\min_{f \in \mathcal{H}} \sum_{i=1}^{n} [y_i - f(x)]^2 + \lambda \int |f''(u)|^2 du, \quad \lambda \geq 0$$

where $f''(x)$ denotes the second derivative and $\mathcal{H}$ is the space of twice differentiable functions. Given the sufficient sequence of observations, consider the extreme cases of smoothing spline fits.

a) (2 marks) When $\lambda \to \infty$, what would the fitted $\hat{f}(x)$ become?

b) (2 marks) When $\lambda \to 0$, what would the fitted $\hat{f}(x)$ become?

Q4. (STAT3612 Hall of Fame)

a) (1mark) Identify (by square) the statistician who promoted the two cultures of statistical modeling. What is his or her name?

b) (1 mark) Identify (by circle) the statistician who had the visionary prediction of the future of data science more than 50 years ago. What is his or her name?