Let $p = 1$ and $\mathcal{D} = \left\{ \begin{bmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{bmatrix}, \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \right\}$ where $\mathcal{Y} \in \mathbb{R}$. We want to create a g for prediction using regression since $\dagger \subset \mathbb{R}$, not $\Big\{0, 1\Big\}$. We want to classify $\hat{y} = g(x^*)$. Let

$$H = \Big\{\vec{w} \cdot \vec{x} : w \in \mathbb{R}^{p+1}\Big\} = \Big\{w_0 + w_1 x : w_0 \in \mathbb{R}, w_1 \in \mathbb{R}\Big\}$$

Let $g = \mathcal{A}(\mathcal{D}, \mathcal{H})$. Find an $\mathcal{A}$ that'll fit the two parameters w_0 and w_1.
Let $\mathcal{A}$ be ordinary least square regression. This requires solving the following problem:

$$\begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \overset{\text{argmin}}{w_0, w_1} \Big\{SSE\Big\} = \sum_{i=1}^{n} (y_i - g(x_i))^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (w_0 + w_1 x_i))^2 = \sum_{i=1}^{n} e_i^2$$

We call these least square estimates b_0 and b_1 and can also be denoted as $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively.
How well does the model predict?

$$SSE = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2 \qquad \text{sum of squared error, units: } y^2$$

$$MSE = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - (b_0 + b_2 x_i))^2 \qquad \text{mean squared error, units: } y^2$$

$$RMSE = \sqrt{MSE} \qquad \text{root mean squared error, units: } y$$

The RMSE expresses the corrected variance of the e's. The RMSE is just like standard deviation and has the same units. It is common to report prediction error as RMSE.
Another well known metric is called R^2, or the proportion of sample variance explained.
Consider the null model. What is the SSE of this model?

$$SSE_0 = \sum (y_i - \overline{y})^2 = (n-1)s_y^2 = SST \text{ sum of squared total}$$

After the model is fit, there is a new, hopefully lower SSE.

$$SSE = \sum e_i^2 = (n-1)s_e^2$$

This means that the model has less error and variance is explained.
How much SSE/s^2 has been reduced as a proportion of the null SSE/s^2?

$$\begin{aligned} R^2 &= \frac{SSE_0 - SSE}{SSE_0} = \frac{s_y^2 - s_e^2}{s_y^2} \\ &= \frac{SST - SSE}{SST} \\ &= 1 - \frac{SSE}{SST} \\ &= SSR \end{aligned}$$

Furthermore, the proportion of sample variance explained is estimated as

$$\frac{s_y^2 - s_e^2}{s_y^2} = \frac{\text{Var}[Y] - \text{Var}[E]}{\text{Var}[Y]}$$

Sine $s^2 > 0$, then $R^2 \leq 1$. Can $R^2 < 0$? Yes. What about $s_e^2 > s_y^2$? This means that the model is worse than the null model
Another way to see this is as follows: Let the null model be $g(x) = \overline{y}$. The residuals will be $e_i = y_i - \overline{y}$. Fit a simple linear regression model $g(x) = b_0 + b_1 x$. Then the residuals will be $e_i = y_i - (b_0 + b_1 x_i)$. This is a much narrower graph because s_e^2 dropped a lot.
RMSE vs. R^2: Which is more important for assessing predictive ability? RMSE. It answers how good the predictions are and the standard deviation of the predictions.
As R^2 increases, RMSE decreases. As R^2 decreases, RMSE increases.
If $R^2 = 99\%$, the RMSE could still be big. Maybe there was a ton of variance in y. You explained most of it but there is still a lot left.

$$RMSE \approx s_e$$

Empirical Rule: $\hat{y} \pm 2 \cdot s_e \approx 95\%$ of all predictions (if $E \sim N(0, \sigma^2)$). Also, $\hat{y} \pm 3 \cdot s_e \approx 99.7\%$ of all predictions.