Solutions

April 5, 2023

Answer Q.1. The data I choose is called Video Game Sales with Ratings which is publicly available on Kaggle. You can find the URL in reference page. There are 6925 observations and 16 variables. This data contains a mixture of categorical variables and numeric variables such as rating (categorical), developer (categorical) and user count (numeric).

In the initial investigation, we find that there are missing values in this data and we choose to ignore them all. There are several variables can be chosen to be response variable such as sales in different regions. However, we choose global sales as our predictor variable and drop the others. The number of predictor variables which are included in the model is also an issue. We will discuss it later. The following figure shows the relationship between global sale and platform.

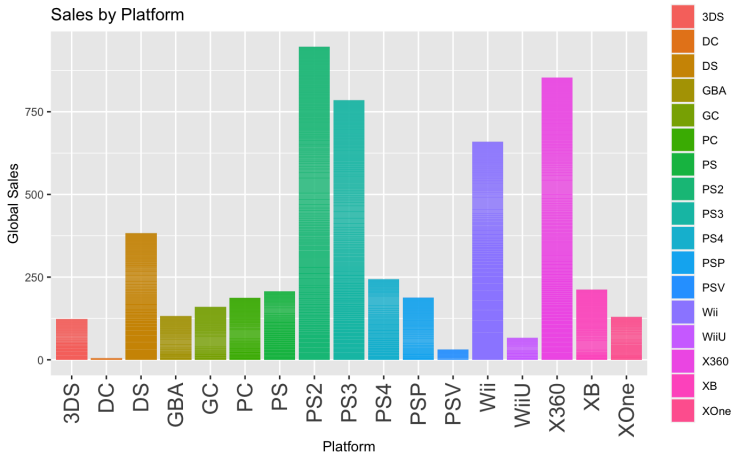

Figure 1: Global sales under different platforms

The model we choose is decision tree which is a basic tree model and we use cross validation to prune the tree so that the decision tree is not bloated. A decision tree does not have a single, specific loss function like some other machine learning algorithms (e.g., linear regression or logistic regression). Since the response variable here is a numeric variable, we will just simply use mean squared error to select the best model. After data cleaning, there are 10 predictor variables in the data and we would expect the decision tree to be very bloated. Thus we only choose some of them. The following figure is a heatmap of correlation.

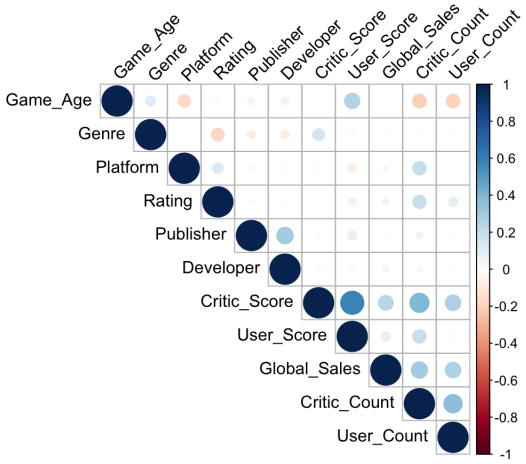

Figure 2: Caption

From the figure above, we can see that most response variables have no correlation with global sale except for critic score, critic count and user count. Thus, we choose these three variables to enter the model. We split our data into training set and test set where 75% is training set and the remaining is test set. The following figures is the decision tree before and after pruning.

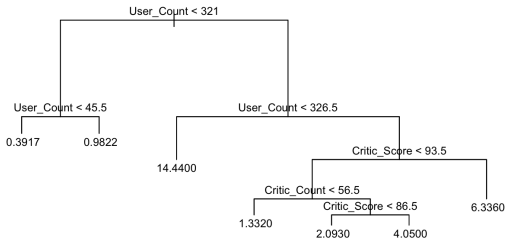

Figure 3: Decision tree

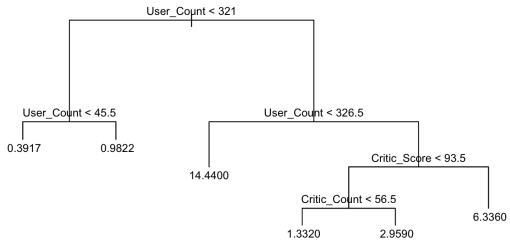

Figure 4: Decision tree after pruning

We don't see too much difference because only two nodes are missing. I would expect a more complex decision tree if we add more variables. However, the mean squared error decreases from 3.216819 to 3.182293 after pruning. Since the result of decision tree is a little pale, we also perform random forest to check the importance of the variables. The error decreases rapidly then the number of trees equals approximately 25 and converge when the number of trees equals 100. From the following figure, we can see the importance of variables, and user count, critic count and critic score are top 3. Since a random forest is an ensemble machine learning algorithm that combines multiple decision trees to create a more accurate and robust model, we usually expect a better performance. The mean squared error in this case it 1.886174 which is almost a half of mean squared error of decision tree.

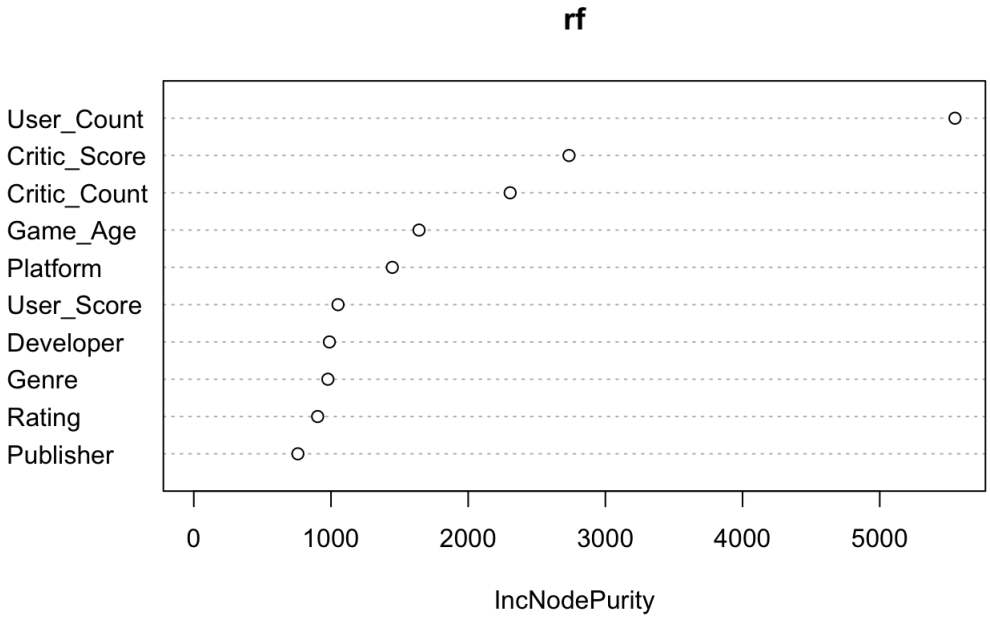

Figure 5: Importance of variables

Answer Q.2. 1 We want to find γ_{jm} such that $\sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma_{jm})$ is minimized. We start with L_2 norm where g is a function of γ_{jm} and

$$\begin{aligned} g(\gamma_{jm}) &= \sum_{x_i \in R_{jm}} [y_i - (f_{m-1}(x_i) + \gamma_{jm})]^2 \\ &= \sum_{x_i \in R_{jm}} [y_i^2 - 2y_i(f_{m-1}(x_i) + \gamma_{jm}) + (f_{m-1}(x_i) + \gamma_{jm})^2] \\ &= \sum_{x_i \in R_{jm}} [y_i^2 - 2y_i f_{m-1}(x_i) - 2y_i\gamma_{jm} + f_{m-1}^2(x_i) + 2f_{m-1}(x_i)\gamma_{jm} + \gamma_{jm}^2] \end{aligned}$$

Take partial derivative with respect to γ_{jm} and set the partial derivative to 0, then we have

$$\begin{aligned} \frac{\partial L}{\partial \gamma_{jm}} &= \sum_{x_i \in R_{jm}} -2y_i + 2f_{m-1}(x_i) + 2\gamma_{jm} = 0 \\ &\implies \gamma_{jm} = \frac{1}{N} \sum_{x_i \in R_{jm}} (y_i - f_{m-1}(x_i)) \end{aligned}$$

2 Next let's have a look at the binomial deviance where g is still a function of γ_{jm} and

$$\begin{aligned} g(\gamma_{jm}) &= \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma_{jm}) \\ &= \sum_{x_i \in R_{jm}} -log(1 + e^{-2y_i(f_{m-1}(x_i)+\gamma_{jm})}) \end{aligned}$$

Take partial derivative with respect to γ_{jm} again and set the partial derivative to 0, then we have

$$\begin{aligned} \frac{\partial g}{\partial \gamma_{jm}} &= \sum_{x_i \in R_{jm}} \frac{1}{1 + e^{-2y_i(f_{m-1}(x_i)+\gamma_{jm}}}(-2y_i)(-e^{-2y_i(f_{m-1}(x_i)+\gamma_{jm}}) = 0 \\ &\implies \sum_{x_i \in R_{jm}} e^{-2y_i(f_{m-1}(x_i)+\gamma_{jm}} y_i = 0 \end{aligned}$$

I don't know whether there is a closed form solution for γ_{jm} in this case.

3 In gradient boosting, we use second order derivative to estimate. We have

$$\sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma_{jm}) = \sum_{x_i \in R_{jm}} [L(y_i, f_{m-1}(x_i)) + \gamma_{jm} g_i + \frac{1}{2}\gamma_{jm} h_i^2] + \Omega(f)$$

where the last term is penalty term and g_i ,h_i^2 are first and second order partial derivative with respect to f, ie.,

$$g_i = \frac{\partial L(y_i, f_{m-1}(x_i))}{\partial f}, g_i^2 = \frac{\partial^2 L(y_i, f_{m-1}(x_i))}{\partial f^2}$$

Take partial derivative at both sides with respect to γ_{jm} we have

$$\begin{aligned} \frac{\partial g}{\partial \gamma_{jm}} &= \sum_{x_i \in R_{jm}} h_i + \gamma_{jm} g_i^2 \\ \implies \gamma_{jm} &= -\sum_{x_i \in R_{jm}} \frac{h_i}{g_i^2} \end{aligned}$$

Thus, for mean squared error, we have

$$\begin{aligned} L &= [y_i - f_{m-1}(x_i)]^2 \\ h_i &= -2(y_i - f_{m-1}(x_i)) \\ g_i^2 &= 2 \\ \implies \gamma_{jm} &= \frac{1}{N} \sum_{x_i \in R_{jm}} (y_i - f_{m-1}(x_i)) \end{aligned}$$

For binomial deviance we have

$$
\begin{aligned}
L &= -log(1 + e^{-2y_i f_{m-1}(x_i)}) \\
h_i &= -\frac{e^{-2y_i f_{m-1}(x_i)}(-2y_i)}{1 + e^{-2y_i f_{m-1}(x_i)}} \\
&= \frac{2y_i e^{-2y_i f_{m-1}(x_i)}}{1 + e^{-2y_i f_{m-1}(x_i)}} \\
g_i^2 &= \frac{-4y_i^2 e^{-2y_i f_{m-1}(x_i)}(1 + e^{-2y_i f_{m-1}(x_i)}) + 2y_i e^{-2y_i f_{m-1}(x_i)} 2y_i e^{-2y_i f_{m-1}(x_i)}}{(1 + e^{-2y_i f_{m-1}(x_i)})^2} \\
&= \frac{-4y_i^2 e^{-2y_i f_{m-1}(x_i)}}{(1 + e^{-2y_i f_{m-1}(x_i)})^2} \\
\implies \gamma_{jm} &= \sum_{x_i \in R_{jm}} \frac{h_i}{g_i^2} \\
&= \sum_{x_i \in R_{jm}} \frac{1 + e^{-2y_i f_{m-1}(x_i)}}{2y_i}
\end{aligned}
$$