Data Science test for Cryoocyte - Brief Intro

Jiayi Ye (Lucy)

Data Preprocessing

First, we want to read in training dataset and preprocess the dataset, For each of the non-numeric variables, we want to see if they are significant to y by simply plotting the means in each variable, If so, then we keep these variables and transform the data type/format. Transform data: categorical data and delete the $ or % symbols.

Second, we want to divide the dataset into Train and Validation set (70%-30%) and rescale the data because of the data scale imbalance.

Third, we start to apply models to the preprocessed dataset, both Training and Validation. Then we evaluate which model has a higher accuracy.

Models

Since the response variable(y) is a continuous variable, the models I chose here are linear regression and lasso regression.

Linear Regression: simplest model for this type of problem. Because of too many x variables, we need to scale them because of the data scale imbalance.

Lasso Regression: It is useful in some contexts due to its tendency to prefer solutions with fewer parameter values, effectively reducing the number of variables upon which the given solution is dependent. Thus, it is a ideal model to solve this high dimensional dataset.

Model Comparison

We use R_square and RMSE to measure the accuracy of the model.

- R-squared (R2) is a measure of the global fit of the model.
- Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors). Residuals are a measure of how far from the regression line data points are; RMSE is a measure of how spread out these residuals are.

The higher the r2 or the lower RMSE, the better accuracy.
We compare these two model using r2 and rmse in the validation set.

	Linear Reg	Lasso Reg
R_2	0.9348168892671871	0.9372997827764117
RMSE	0.051469210482093235	0.050479435294885425

Both of the regressions has high R_2 and small RMSE, while Lasso Regression is slightly better then the Linear Regression. Then I chose to use Lasso Regression to predict y for the test dataset.