# Assignment 1 Machine Learning Models

## Materials

Here is the Google colab link for this assignment: <https://drive.google.com/file/d/1vCXyvXrPqxh11NYX_CHYEvJLPkrLyhs5/view?usp=sharing>

public dataset link: <https://www.kaggle.com/lava18/google-play-store-apps>

The license: This work is licensed under the Creative Commons Attribution 3.0 Unported License. To view a copy of this license, visit <http://creativecommons.org/licenses/by/3.0/>.

## Idea & abstract

This dataset contains the details of applications in Google Play Store (categories, rating, size, download times, reviews, etc.). I hope to build several prediction models to predict the rating for applications with Linear Regression, Random Forest Regression and Support Vector Machine.

## Preprocessing and approaches

Because this dataset contains lots of meaningless sample, so I had to do data cleaning before we got started(The steps of cleaning are in the .ipynb file).

I tried to figure out the relationship between rating of applications and their reviews, size, install numbers, types, prices, categories. To prevent overfitting models, I used cross-validation,

trained with more data, and removed some irrelevant features based on reality experience. I used the mean square error to evaluate the performance of model. I think this metric really makes sense to display the deviation between predicted result and real data.

## Conclusion

According to the mean square error, it’s hard to conclude which model has the best predictive accuracy and lowest error. Compared with my models based on the mean error, I think the Random Forest Regression model has the lowest overall error rates. And SVM and linear regression model are very close. But I don’t think there is a strong relationship between rating and other attributes from dataset. There is no data displaying the quality of application in Google Play Store except rating. Even though the number of reviews states the popularity of a specific application, sometimes the reality might get polarization. So I think there is no actual hyperparameter in this dataset.

## Q&A

\* Is the relationship significant?

I don’t think the relationships in my models are very significant. However, there are existing some relationships. Especially I don’t think there is linear relationship between rating from users and other attributes.

\* Are any model assumptions violated?

No. There is no model assumption violation.

\* Is there any multicollinearity in the model?

No. There is no multicollinearity in the model. I tested it with Variance Inflation Factors to identify correlations between variables.

\* In the multivariate models are predictor variables independent of all the other predictor variables?

The predictor variable is absolutely independent.

\* In multivariate models rank the most significant predictor variables and exclude insignificant ones from the model?

Sorry, in my task I just had only one predictor variables. And I had removed some irrelevant features from dataset.

\* Does the model make sense?

I don’t think these models totally make sense. Apparently, the linear regression model doesn’t fit well.

\* Does regularization help?

Yes. Regularization is working in logistic regression model.

\* Which independent variables are significant?

Reviews, install numbers, prices, and size.

\* Which hyperparameters are important?

Reviews and install numbers are important.