**Introduction**

To Develop Model that can identify the similarities or dissimilarities of Dublin City and Paris City

**Problem Statement**

Both Cities are best attraction for tourism. Every year thousands of tourists visiting both places. Both cities have cafes, restaurant, hotels, side views and so on. So, it is very difficult to identify with is the best option to LIVE. Place where you can work and live your life? Which city is expensive for living and food? Which City is safe to live for long run? To identify similarities or dissimilarities of Dublin City and Paris City.

**Data**

I will be going to use foursquare data set for analysis. Foursquare Data will tell me about similarities or dissimilarities of Dublin City and Paris City. Apart from that I will use crime data from both the cities and which city is best optimum for long run.

**Process Methodology and Result**

Predictive Modeling There are two types of models, regression and classification, that can be used to predict player improvement. Regression models can provide additional information on the amount of improvement, while classification models focus on the probabilities a player might improve. The underlying algorithms are similar between regression and classification models, but different audience might prefer one over the other. For example, an NBA team executive might be more interested in the amount of improvement (regression models), but a general NBA fan might find the results of classification models more interpretable. Therefore, in this study, I carried out both regression and classification modeling. 4.1 Regression models 4.1.1 Applying standard algorithms and their problems I applied linear models (linear regression, Ridge regression, and Lasso regression), support vector machines (SVM), random forest, and gradient boost models to the dataset, using root mean squared error (RMSE) as the tuning and evaluation metric. The results all had the same problems. The predicted values had much narrow range than the actual values (Figure 8), and as a result, the prediction errors were larger as the actual values deviated further from zero (Figure 9). These results were not acceptable, because players with large improvement/decline were arguably more important for NBA teams to predict than players with little change in performance. Having larger errors on those predictions was obviously not desirable. 4.1.2

**Discussion**

The reason behind these problems were the uneven distribution of player improvement, in that players with little improvement/decline were more common than players with big improvement/decline (Figure 8). Therefore, the models tried to prioritize minimizing errors on players with little improvement/decline when RMSE was used as the evaluation metric. My solution to this problem was to assign weights to samples based on the inverse of the abundances of target values. In other words, players with large improvement/decline would have higher weights in model training and evaluation because they were more rare. Using this method, all models predicted target values with similar range and distribution as the actual target values (Figure 10). Figure 8. Distribution of actual and predicted improvement using linear regression with equal weights of samples. Figure 9. Scatterplot of prediction errors vs. actual target values using linear regression with equal weights of samples. Figure 10. Distribution of actual and predicted improvement using linear regression with different weights of samples based on inverse of sample abundance.

**Conclusion**

Performances of different models Using the new approach of different sample weights, I built linear regression, SVM, random forest, and gradient boost models using weighted root mean squared error as the evaluation metric. For each model, hyperparameters were tuned using the same metric and cross validation. For comparison, I also built a simple linear regression model with just one independent variable (age) as the benchmark model. SVM had the best performance among all models, which had ~26% less error than the benchmark model (Table 2). The predicted improvements had linear relationship with the actual improvements (Figure 11). Table 2. Performance of the regression models. Benchmark (one feature) Linear Regression SVM Random Forest Gradient Boost Weighted RMSE 3.84 2.98 2.86 2.93 2.96 4.2 Classification models The application of classification models was much more straightforward. I divided the samples into two classes (improvement>=0 or