**FINAL PROJECT WRITEUP**

Our group comprises of Aman Geetey (116087599) and Pranav P. Jaipurkar (115798120). The dataset which we used for our project was the world university rankings dataset taken from <https://www.kaggle.com/mylesoneill/world-university-rankings>. The dataset had files named cwur, times, shanghai, education expenditure and school and country table. Major portion of our project with data analysis on data from files cwur, times and shanghai.

In the initial stage, Pranav did some basic exploration of the data across all files. For example, he found the count of total number of NA values in all files, total number of countries from which universities were recorded along with its summary, basic statistical quantities such as mean of score in cwur dataset across USA, UK and India, range of the year through which the data was present, range of ranks allotted to universities in different year. Thereafter, he made some visualization plots using “ggplot2” package in R, one out of which showed that there is an inverse exponential relationship with score and world rank. Using “dyplr” package, Aman found out the total number of distinct institutions and countries present in the cwur dataset. He did piping to find the median of alumni employment, citations and world rank. Using “stringr” package, Aman replaced some of the strings like university, Massachusetts Institute of Technology, California, Michigan, United Kingdom, South Korea and Federal Institute of Technology with shorter strings. In cwur dataset, Pranav found out correlation between all the numeric variables was found out using 3 different methods named pearson, spearman and kendall. This correlation was plotted using “corrplot” library. The above process gave us an introductory idea about which factor dominates in predicting world rank.

Pranav made a linear regression model to predict the world rank using research components like publications, influence, citations patents and employment-based models using components of alumni employment, quality of education. Thereafter, predictions were made from this models and accuracy was tested which was significantly low.

Aman made a subset of times dataset with universities belonging only to USA. We used “dyplr” package to see the universities in descending world rank. Correlation between all the numeric components in the times\_sub dataset was found out and correlation plots were made.

Pranav: Now, one of the main idea of our project was to use logistic regression models to find whether a university is a top ranked university or not. To complete the task, I divided the complete cwur dataset into 4 parts according to the year. Now, the year 2012 and 2013 had data only for 200 universities which was not sufficient to build logistic regression models and test their accuracy. Hence, I made logistic regression models for complete cwur dataset and for data from years 2014 & 2015 in cwur dataset. For all the models, I divided the data into training (70%) and testing (30%). The accuracy of all the models was tested on the respective test datasets. The accuracy of logistic regression model for complete cwur dataset taking into account the employment factors was around 90% whereas it was around 92% if research component variables were taken into consideration. This meant that research components are slightly more dominant in predicting rank than employment components. When the logistic regression models were made for cwur 2014 dataset, then the accuracy of the logistic regression model taking into consideration the research factors was around 91%, whereas it was around 94% for logistic regression model made from cwur 2015 dataset and taking employment factors into consideration. Hence, it could be concluded that if models are made from dataset comprising data for a particular year, then the accuracy of that model would be optimum.

After looking at the exceptionally high accuracy of the logistic regression models made above, I thought of making separate logistic regression models for a country according to year 2014 and 2015. In this process, I made a separate object which had data for the USA universities for year 2014 and 2015. A separate column of top\_rank was allotted to this object and if the national rank of the universities was greater than 40 then 0 was allotted otherwise 1 was allotted. The employment factors based logistic regression model for USA universities for the year 2014 had an accuracy of around 85%, whereas research based logistic regression model for USA universities for the year 2014 had an accuracy of about 87.5%. As always, the accuracy of the research-based model was greater than employment-based model but the overall accuracy was decreased as compared to previous models. I think that the main reason for this was the lack of observations in the training and testing dataset. The accuracy of employment based logistic regression model for USA universities for year 2015 was around 84%, whereas it was around 82% for research-based models.

Now moving from cwur dataset to times dataset. The same logic was applied to times dataset. A new column of top\_rank was allotted to the times dataset where 1 was allotted if the rank was less than 150 and 0 was allotted otherwise. After that a logistic regression model was made taking into account all the variables which had numerical data type. The accuracy of this logistic regression model was tested on test data and was found to be around 72%. Thereafter, I made 2 logistic regression models based on research parameters and employment parameters. The accuracy of employment-based model was around 70% whereas it was around 67% for research based models. Hence, I concluded that times magazine places more emphasis on employment rather than research for rank determination.

Lastly, the same logic was applied to shanghai dataset. A new column named top\_rank was allotted to the dataset, where the column had a value 1 if rank of the university was less than 100 and 0 otherwise. The accuracy of the logistic regression model made by taking into account all the variables having a numerical quantity was around 45% which was significantly less. I think that the main reason for this would be that shanghai magazine uses a very complex relationship of all the independent variables to determine the world rank.

I think the idea of using logistic regression models to predict whether a university is of a given standard or not was a pretty decent idea as the accuracy of all the models made above was high (above 87%) for maximum models. Also, if I would be given information regarding the independent variable of the models for the future years, then the models would pretty decently classify whether the university is above our given standards or not. Hence, I think this technique worked as per the criteria set.

Aman: Ordered ranking systems are heavily reliant on a host of data that is being provided to them to predict the best in class and the poorest performer in the spectrum of all the choices at the disposal of the end user. Exactly this requirement is at the heart of all such models where the end goal is to predict a rank of each individual member in the given dataset, like in our dataset. Oftentimes, in these models, certain features play a more imperative role than others. For example, a certain university ranking system could rank an obscure university at the top of the leaderboard and another Ivy league just below it. The ranking systems have different ranking algorithms at their disposal, hence all the ranking have oftentimes major but sometimes slight differences in the ranking of same universities. This can be verified from the comparison of two prominent ranking system for universities, namely the QS and Times Higher Education ranking systems. The top 50 rankings for some years had considerable differences.

I tried to incorporate a similar model in his case which can be generalized further for the prediction of the strategy that is adopted by most of the university ranking systems. The problem statement can be generalized as follows-

“*Given a dataset we need to predict what features play a dominant role in deciding the final score for classifying the universities and ranking them.”*

The general approach to go about solving this kind of problems is to assume that each system has an algorithm in place which assigns certain weights to the feature, applying a larger weight to the features that are more important. For example, a certain higher education ranking system might be more biased towards the total research output generated by a certain university. Hence, it would apply a larger weight to the features that represent the research output of that university.

My hypothesis was that if we build a classification model based on a score that is the weighted sum of some features of the dataset, and then build models using existing algorithms to predict which features could have larger weights compared to others. I took the cwur dataset for this purpose because of the wealth of features it contains specially for research, since the classification methodology I have used focusses more on research. This analysis could provide us with the distinguishing factors when we are comparing two ranking systems, for example one might consider quality of education to be a factor which should have greater importance, while other might consider only the number of citations to be the deciding factor.

After defining the consolidated feature, can be seen in the dataset as gen\_res, he applied SVM models trained with different features. The classification in general gave good results, most giving an accuracy more than 80 percent. However, the maximum accuracy (99%) was reached when we trained the models with the features the actual model was made out of. Thus, I have at least now isolated the feature which definitely carry a lot of weight. Hence, I think that this technique worked well.