**BIA686 – Final Analysis Project Brief - KKBox's Churn Prediction Challenge**

**Sin Cheng, Olaoluwa Thomas & Thushara Tom (Fall 2017)**

WSDM KKBox’s Churn Prediction Challenge is a competition hosted on Kaggle to predict if the user will churn after their subscription ends. There are five datasets used for the analysis – train dataset, transaction dataset, members dataset, user logs dataset and the final submissions file. The train data contains user ids and whether the user churned or not. The final submissions dataset consists of the user ids we are supposed to predict the churn for. We ran the datasets on Alteryx for data pre-processing to create new features and combine the datasets. A train and test dataset were created consisting of features associated to the users. The train dataset consists of 970,960 records and 147 features. The missing values were imputed using mode for categorical features and median for continuous features. 9% of our dataset represents churn in the train dataset. We ran random forest classifier model on the dataset. The most significant features from the model are: list\_price\_1599 (users with plan worth $1599), no\_of\_cancellations (the number of cancellations an user made), registered\_via (the registration method of user), gender (male or female) and payment\_method\_36 (users that use payment method 36). Predictions were made on the test dataset and a submission was made to Kaggle on 29th November 2017. As of today, we rank as 185 out of 350 with a score of 0.1552.