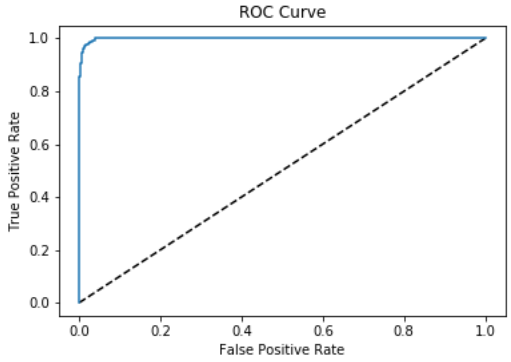
Relax Inc Take Home Data Science Challenge

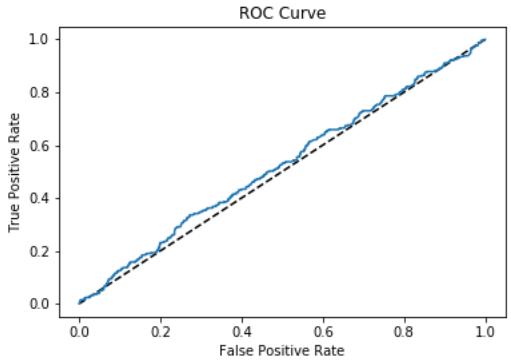
The problem statement here is to identify the user adoption given a set of features. For this purpose, the data was cleaned up to be represented in a dataframe consisting of the following features:

1. opted_in_to_mailing_list
2. enabled_for_marketing_drip
3. org_id
4. visit_count
5. source_ORG_INVITE
6. source_PERSONAL_PROJECTS
7. source_SIGNUP
8. source_SIGNUP_GOOGLE_AUTH
9. user_invite

The adopted class was calculated using the provided condition of 3 logins on separate days. Following this, a Random forest based model was created which resulted in a high AROC - curve shown below:



On further investigation on feature importance - it was noted that the "visit_count" was the feature which overpowered all other features. In a way it makes sense since usage history will be a strong indicator of user adoption. However, if adoption probability of a "new" user is of interest, then this feature will have to be dropped. This was done and in that case, the ROC curve was not much better than a random prediction:



In conclusion, it can be said that without usage history it is very difficult to predict future user "adoption". However, if there is some usage history data, it will be very easy to predict very accurately which users will and will not be adopted.