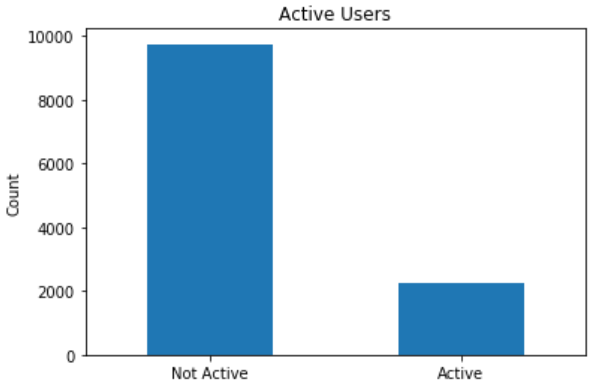
Relax Challenge

After the first few steps of loading, we need to label users as active or not according to the criteria of creating 3 sessions within a week, with each session being at least 1 day apart. To be labelled as active a user must have done this at least once within the engagement period.



The data is slightly imbalanced, suggesting we should look at precision, recall, and f1 scores as metrics for our model in addition to accuracy.

Once we have each user labelled as active or not, we can also create some other features for our model to learn from. Let's include in our features a column indicating the number of days between a users last created session, and their creation date. Let's also include a column indicating the number of sessions a user has created based on the number of entries a user_id has in the user engagement data.

With the data we have wrangled, we can now implement a basic logistic regression model.

accuracy: 0.999 precision: 1.000 recall: 0.996 f1 score: 0.998

Our model looks great! Now from this model we can look at the coefficients it used for each feature.

	feature	coef
0	sessions	3.003008
1	number_of_days	0.025632
2	invited_by_user_id	-0.000141

We can see that the number of sessions that a user has created is the best indicator of whether or not a user is active.