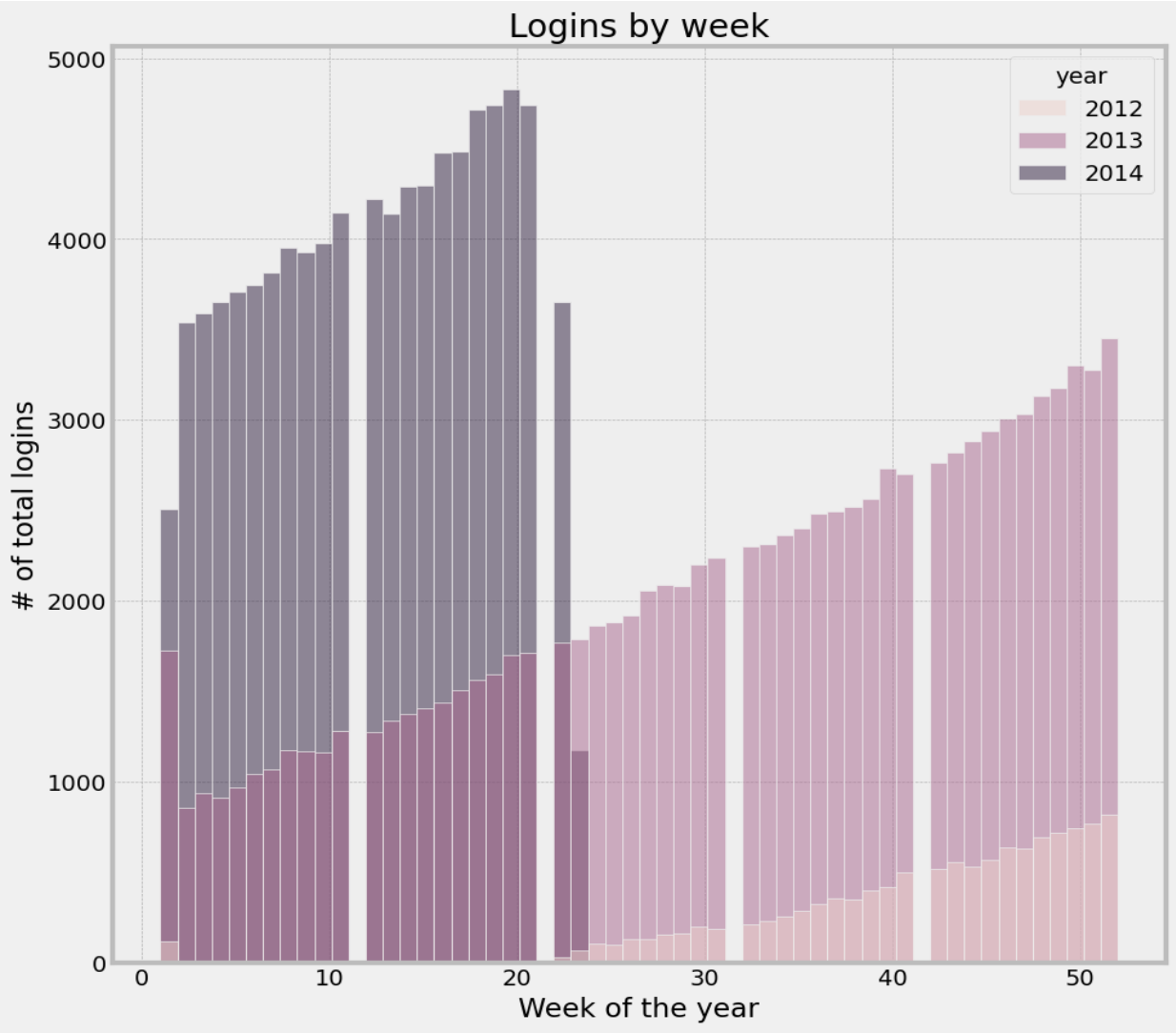
Relax Data Science Challenge

This project was conducted to answer the problem question: What factors best predict future user adoption of the product?

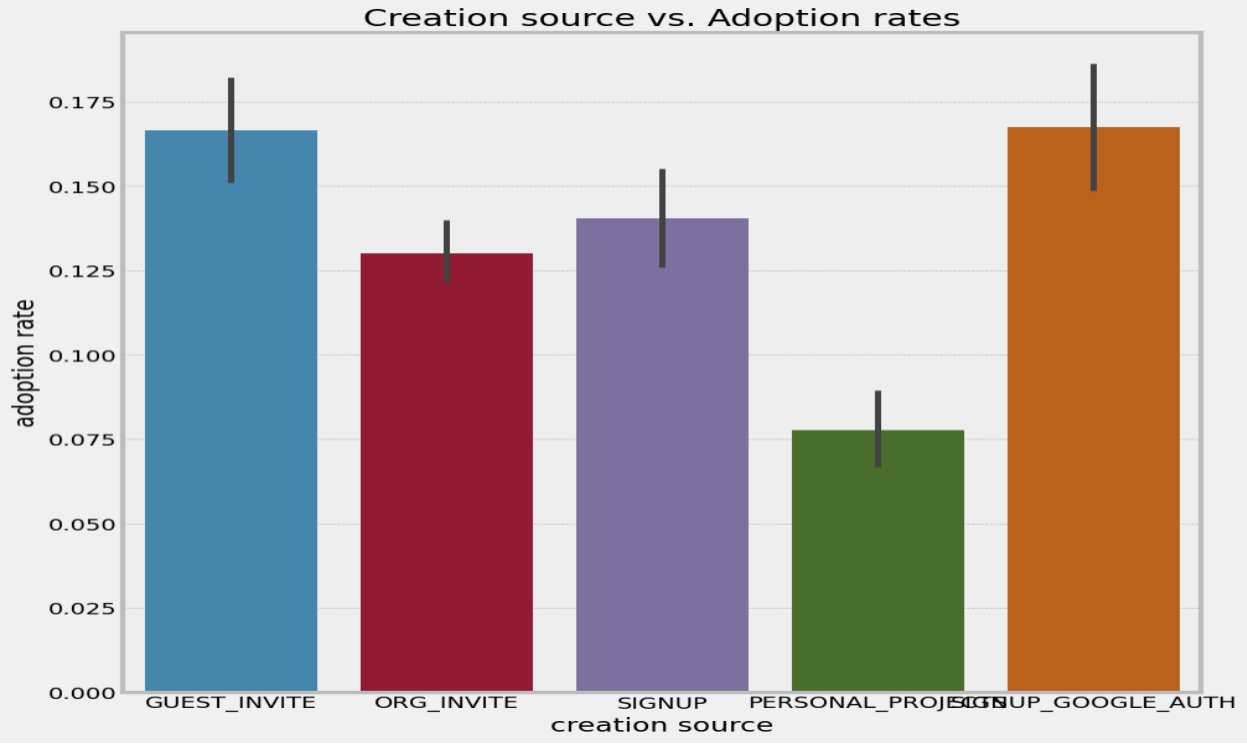
Data Wrangling:

Data from the logins table was processed to find users who had logged in at least 3 times in any 7-day period and defined them as an 'adopted user'. This list of users was then merged with the users table and a target variable was created to define each user as adopted or not.

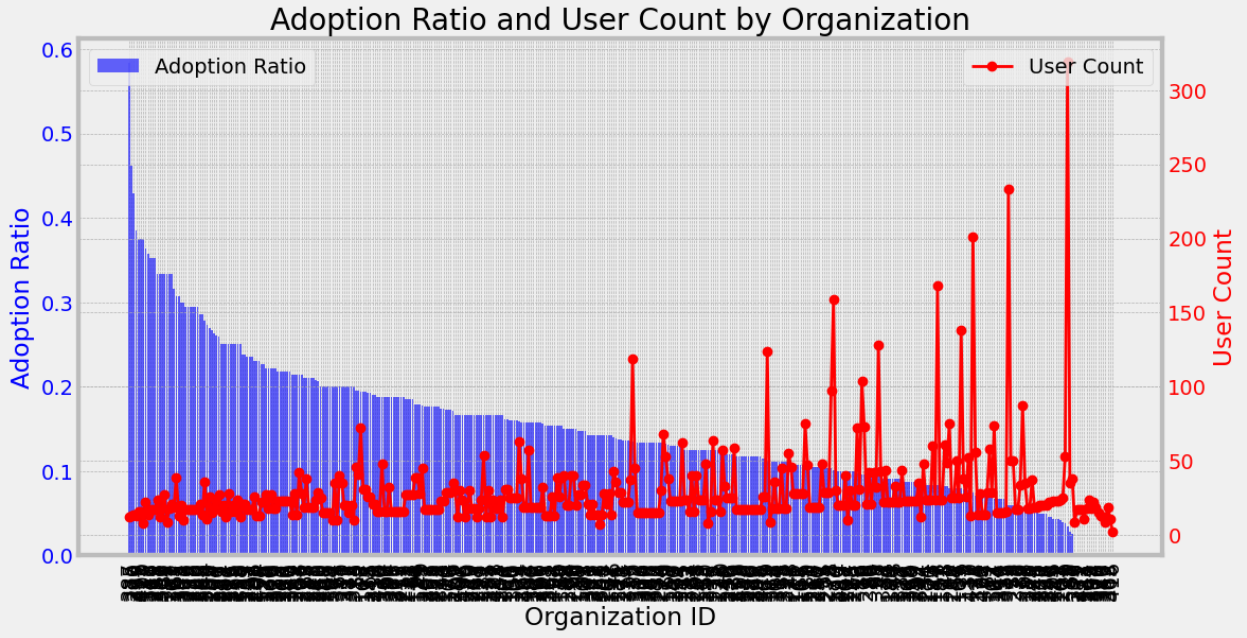
Exploratory Data Analysis:



Data shows that total logins have been steadily increasing over time, year after year.



When looking at account creation sources, personal projects seem to have significantly lower adoption rate than other sources.



Looking at organization id's showed that organizations with higher count of users tended to have lower adoption rates.

Modeling:

Multiple models were tested to predict user adoption. Prior to modeling, the data was pre-processed by converting categorical variables into binary columns and removing unnecessary columns. A column showing total logins per user was removed because it was so closely tied to the definition of an adopted user to prevent circular logic in the model.

Training and testing splits were created. Logistic Regression model performed very well, with a high accuracy score (0.97) and high precision and recall. The model showed that the most important features were whether the user opted in to the mailing list, creation source and organization id.

A random forest model was also tested. While it performed well (accuracy score = 0.96), it had much lower recall for the adopted user class. Also, looking at its important features, it was heavily indexed on days since last login and days since creation.

Summary:

The logistic regression model does an excellent job of predicting future user adoption with the given data and with low complexity cost. It also shows that opting into mailing list, creation source and organization are the most important features in terms of predicting user adoption. Exploratory analysis showed that the marketing drip did not appear to have significant difference in user adoption.

Future Work/Additional Info/Considerations

To further improve the model, more user data can be collected. Also, additional feature engineering with more demographic data on users including usage time per login or even user ratings could further help the model.