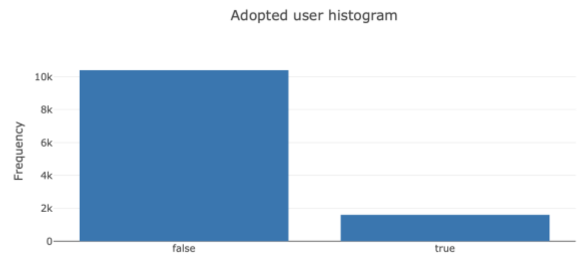
Driving factors behind Asana's user adoption

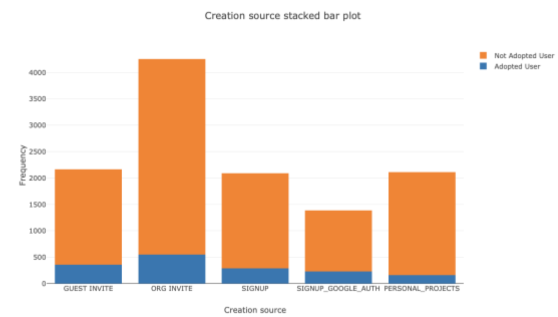
In this problem, our aim is to predict the driving factors behind asana's user adoption. We define an "adopted user" as a user who has logged into the product on three separate days in at least one seven-day period. Since adopted users are more likely to be successful at using Asana in the long term than those that are not adopted, we want to know what things are likely indicators of future adoption. With this in mind, we'd like to identify which factors that predict user adoption. With the intention to understand the features that contribute to user adoption, descriptive analysis methods are emphasized on analyzing feature importance. The “users” data-set has information of 12,000 users and “user_engagement” data-set provides us with the login activity for all those users over the last few years.

1. Data pre-processing

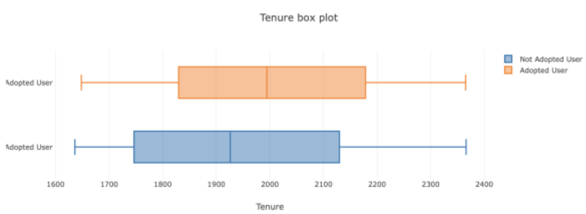
1.1 Target Creation

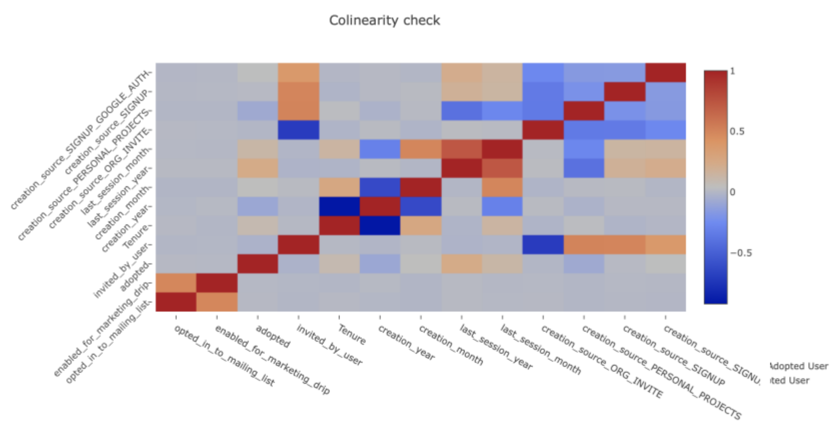
Using the adopted user logic provided for the challenge, we go ahead and create the target feature "adopted" and add it to the “users” data-set. 1602 of the 12,000 users are marked as adopted users i.e. only 14% of the users are marked as an adopted user. Due to the high skewness, we will sample the dataset before modelling.





1.2 Exploration

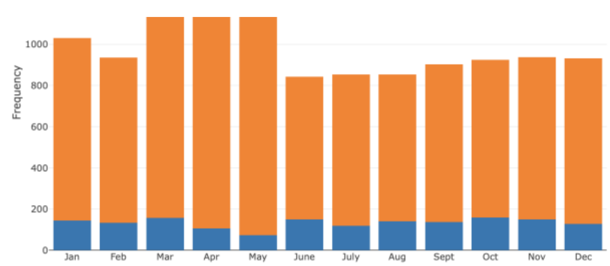




From the plots, we can see that the main creation source is Organization Invite and you can note that adopted users have a higher tenure than non-adopted users. The collinearity check shows us that there is high positive correlation between, last_session_month and last_session_year. Negative correlation between creation_year and tenure which makes sense.

1.3 Feature Engineering

We will create new features like creation_year, creation_month, last_session_month, last_session_year to determine seasonal effects. Also creates field tenure to determine how long-term customer life cycle value affects user adoption. The stacked bar plot against the frequency for creation month shows us that user-adoption is the least in May.



2. Modeling

We will sample the 12,000-entry data-set using the SMOTE (Synthetic Minority Over-sampling Technique) method in-order to make the class variable 'adopted' balanced.

The RandomForestClassifier achieved an accuracy of 94.8%. The feature importance attribute of the RandomForrestClassifier shows us that last_session_year, tenure of the customer, creation_source_ORG_INVITE, invited_by_user and creation_month are the 5 most important factors that affect user adoption. One drawback of the Random Forest is that it lacks interpretation. Next, we would like to know the standard errors related to these features and the effect they have to user adoption in Asana. For this we run the Logistic regression model.

RandomForestClassifier

```
print(confusion_matrix(y_test,y_pred))
print(classification_report(y_test,y_pred))
print(accuracy_score(y_test, y_pred))
```

```
[[2052  102]
 [ 121 2088]]
             precision    recall  f1-score   support

      False       0.94      0.95      0.95      2154
       True       0.95      0.95      0.95      2209

avg / total       0.95      0.95      0.95      4363

0.948888379555
```

Logistic Regression

```
Optimization terminated successfully.
         Current function value: 0.599375
         Iterations 22
```

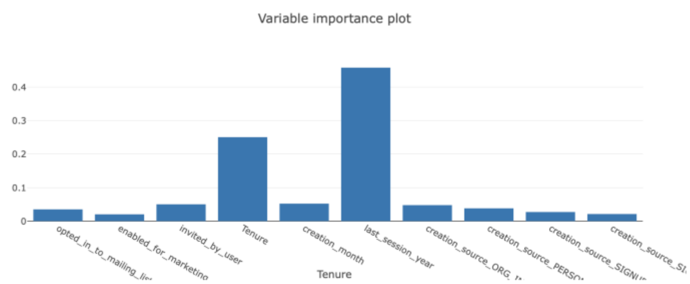
Model:	Logit	Pseudo R-squared:	0.135
Dependent Variable:	adopted	AIC:	17450.2257
Date:	2018-11-21 23:45	BIC:	17518.4888
No. Observations:	14542	Log-Likelihood:	-8716.1
Df Model:	8	LL-Null:	-10080.
Df Residuals:	14533	LLR p-value:	0.0000
Converged:	1.0000	Scale:	1.0000
No. Iterations:	22.0000		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
opted_in_to_mailing_list	-0.0875	0.0524	-1.6709	0.0947	-0.1901	0.0151
enabled_for_marketing_drip	-0.0216	0.0637	-0.3393	0.7344	-0.1464	0.1032
invited_by_user	-0.6607	1473819.3335	-0.0000	1.0000	-2888633.4740	2888632.1526
Tenure	-0.0011	0.0000	-23.3520	0.0000	-0.0012	-0.0010
creation_month	0.0193	0.0054	3.5985	0.0003	0.0088	0.0299
last_session_year	0.0015	0.0000	36.5972	0.0000	0.0014	0.0016
creation_source_ORG_INVITE	-0.7407	0.0551	-13.4454	0.0000	-0.8487	-0.6327
creation_source_PERSONAL_PROJECTS	-0.1955	1473819.3335	-0.0000	1.0000	-2888633.0088	2888632.6178
creation_source_SIGNUP	-0.2640	1473819.3335	-0.0000	1.0000	-2888633.0774	2888632.5493
creation_source_SIGNUP_GOOGLE_AUTH	-0.2011	1473819.3335	-0.0000	1.0000	-2888633.0145	2888632.6122

The regression model shows us that out of the top 5 features from the earlier model, all except invited_by_user has low desired p-value and low standard error. So, we can safely conclude that last_session_year, tenure, creation_source_ORG_INVITE and creation month are important reliable features that drives user adoption.

3. Inferences

Looking into the coef. values from the Logistic regression we can have some takeaways -



1. **Active customers have higher odds of being an adopted user** - Positive coef for last_session_year shows that as last_session_year increases, users who logged in recently have a higher odds of being an adopted user.
2. **Newer customers have higher odds of being and adopted user** - Negative coef for tenure shows that as tenure decreases, there is higher odds of being an adopted user. This would also indicate that as the customer life span increases odds of converting the customer to an adopted user would decrease.
3. **Seasonal effects play a part** - This is indicated by the earlier bar graph and by the model. With a significant decrease in April and May. The model shows that as we go from Jan to Dec, the odds of being an adopted user increase. As the worst months April and May falls in the first half of the year, it does go in sync with the bar graph.
4. **Organization Invites are not very effective** - Strong negative coef for creation_source_ORG_INVITE shows that when it's an org invite, there is lower odds for the user to be an adopted user. This would be an indicator that organization invites are not an effective method to increase adopted users.

In future, I recommend a deep dive into the models to see multi variable interactions. This would help us to determine combined effects of features. For example, how user adoption changes when both tenure and org invites are considered together. It could give us a deeper understanding if there exists a subset of customers (new, medium, or old) which has interactions with org invites to give positive odds on user adoption.