Splash Learn

Problem Statement:

Product and marketing team wants to understand the churn better so that they can identify a user segment for reducing churn. They are also looking for triggers (triggers related to email or content usage) which may lead to churning of users. Product & marketing team expects you to build a churn model that predicts churn and variable importance.

Data Exploration:

User Attributes - This file contain user's subscription and its meta data
Email Data - This file contains when an email is sent to users and what is the category of email
Content Usage - What users have played on a day to day basis?

Restructuring the data

S**o,** we have data from three attributes: user ,email and content. We have to consolidate the data so that we have all the valuable information from three sources to analyse the churn of a particular user.

Feature Engineering

Creating new feature column out of Email data and content usage so that the model can easily find pattern out of the data to better predict churn.
Feature created out of email data are as follows:

email_category_activation_count,email_category_app_monthly_purchase_15day_count, email_category_app_yearly_purchase_135day_count, email_category_app_yearly_purchase_200day_count, email_category_app_yearly_purchase_260day_count, email_category_app_yearly_purchase_320day_count, email_category_app_yearly_purchase_75day_count, email_category_apps_day3_pa_1_count, email_category_apps_day3_pa_2_count, email_category_cancellation_confirmation_count etc.

Similarly, we will be creating new feature from content usage. Some of the feature created out of content usage are as follows:
'activity_asked_count', 'books_asked_count',
'ela_lp_asked_count', 'math_learning_games_asked_count',

'math_lp_asked_count', 'mathfacts_asked_count', 'others_asked_count', 'playzone_asked_count', 'reading_asked_count', 'spelling_asked_count', 'activity_time_spent_secs', 'books_time_spent_secs', 'ela_lp_time_spent_secs', 'math_learning_games_time_spent_secs', 'math_lp_time_spent_secs', 'mathfacts_time_spent_secs', 'others_time_spent_secs', 'playzone_time_spent_secs', 'reading_time_spent_secs', 'spelling_time_spent_secs', 'activity_unique_playables_attempted', 'books_unique_playables_attempted', 'ela_lp_unique_playables_attempted', 'math_learning_games_unique_playables_attempted', 'played_date_max'

Final Merge data

Int64Index: 3138 entries, 0 to 3137
Columns: 121 entries, Subscription ID to played_date_max
dtypes: datetime64[ns](3), float64(40), int64(74), object(4)

Out of these, the data of the data type datetime64[ns] are 'cancellation date', 'played_date_max' are of no use as cancellation date are null for most of the user and played_date_max is the last date when the content was played.

We can extract time related info from the subscription date for example month in which the subscription was bought.
So the final data feature columns now becomes 119

Missing values Analysis

		Missing_Value_%
Feature		
0	Subscription ID	0.000000
1	plan type	0.000000
2	student grade	0.000000
3	acquisition type	0.000000
4	is churned	0.000000

...	...	...
114	others_unique_playables_completed_attempted	0.046526
115	playzone_unique_playables_completed_attempted	0.046526
116	reading_unique_playables_completed_attempted	0.046526
117	spelling_unique_playables_completed_attempted	0.046526
118	subcription_month	0.000000

We can drop the columns with missing values as there are many columns are most of the feature having missing values are irrelevant to churn prediction. This will also help to reduce
Dimension of the data.

Data Imbalance

As we have a target variable i.e. “is churned”. If we look into the count of 0 label and 1 label we can see that the data is imbalanced.

0 0.644719
1 0.355281
Name: is churned, dtype: float64

One hot Encoding

So we have data which categorical in nature, we have used one hot encoding for data transformation.

Model Selection and Model Building

After data pre-processing and EDA, we finally arrived at model building. Since the data is imbalanced and to extract complex data patterns we have used tree based ML algorithms for classification: Algorithms used along with the accuracy are as follows:

Decision Tree Accuracy: 0.7045075125208681
Random Forest Accuracy: 0.7629382303839732
XGBoost Accuracy: 0.7879799666110183

Feature Importance are as follows:

```
email_category_activation_count: 0.04123539232053419
email_category_app_monthly_purchase_15day_count: 0.02520868113522534
email_category_app_yearly_purchase_135day_count: 0.0
email_category_app_yearly_purchase_200day_count: 0.0
email_category_app_yearly_purchase_260day_count: 0.0
email_category_app_yearly_purchase_320day_count: 0.0
email_category_app_yearly_purchase_75day_count: 0.0
email_category_apps_day3_pa_1_count: 0.0
email_category_apps_day3_pa_2_count: 0.0
email_category_cancellation_confirmation_count: 0.0005008347245408995
email_category_clevertap_count: 0.0
email_category_courses_d8_email_free_users_count:
-0.00016694490818031094
email_category_deletionrequestmail_count: 0.0056761268781301945
email_category_inactive_for_7_days_count: 0.0
email_category_live_class_daily_reminder_count: 0.0
email_category_live_class_one_hour_reminder_count: 0.0
email_category_live_class_weekly_reminder_count: 0.0
email_category_live_class_welcome_email_count: 0.029382303839732858
email_category_login otp_count: 0.0
email_category_mid_week_reminder_count: 0.0
email_category_new_math_facts_count: 0.0
email_category_nps_followup_detractor_count: 0.03055091819699496
email_category_nps_followup_passive_count: 0.0
email_category_nps_followup_promoter_count: 0.0
email_category_parentappnotification_count: 0.0
email_category_parentfollowupemail_count: 0.0
email_category_practice_reminder_end_count: 0.0
email_category_practice_reminder_first_count: 0.0
email_category_practice_reminder_mid_count: 0.0
email_category_ptl_assignmentnotificationtonotlinkedparent_count: 0.0
email_category_ptl_teacherinvitesparent_count: 0.0
email_category_ptl_teacherinvitesparent_o_count: 0.0
email_category_ptl_teacherinvitesparent_reminder_count: 0.0
email_category_purchase_receipt_count: 0.0
email_category_push_parent_to_start_trial_count: 0.004006677796327196
email_category_remindpascode_count: 0.0
email_category_socialprivacyemail_count: -0.0005008347245409106
email_category_splashlearn courses reachouts, splashcourses_e1_count:
0.0
email_category_subs_onboarding_day0_count: 0.003839732888146896
email_category_subs_onboarding_day1_count: 0.0
email_category_subs_onboarding_day2_count: 0.01001669449081799
email_category_subs_onboarding_day3_count: 0.0
email_category_subs_onboarding_day4_disappeared_count: 0.0
email_category_subs_onboarding_day4_played_count: 0.0
email_category_unconfirmed email account deletion warning_count: 0.0
email_category_upgrade_monthly_plan_count: 0.0
```

```
email_category_web_day2_count: 0.0
email_category_web_day4_count: 0.0
email_category_web_day7_count: 0.0028380634390650973
email_category_web_monthly_purchase_15day_count: 0.0
email_category_web_preview_end_count: 0.02787979966611014
email_category_web_quarterly_purchase_15day_count: 0.002671118530884775
email_category_web_quarterly_purchase_75day_count: 0.0
email_category_web_yearly_purchase_135day_count: -0.00333889816360603
email_category_web_yearly_purchase_15day_count: 0.0
email_category_web_yearly_purchase_200day_count: 0.0
email_category_web_yearly_purchase_260day_count: 0.0
email_category_web_yearly_purchase_320day_count: 0.0
email_category_web_yearly_purchase_384day_count: 0.0
email_category_web_yearly_purchase_75day_count: 0.0
email_category_webtrialabouttoend_count: 0.0
email_category_week_start_reminder_count: 0.0
email_category_weeklyreport_count: 0.0
email_category_ws_reminder_01a_count: 0.0
email_category_ws_reminder_01b_count: 0.0
email_category_ws_reminder_01c_count: 0.0
email_category_ws_reminder_02a_count: 0.001001669449081799
email_category_ws_reminder_02b_count: 0.0
```

So, these are some of the features along with the feature importance derived from the data. Now once we have feature importance for the features we now have to collaborate with the product and marketing team to decide the relevant feature and retrain the model which is an iterative process.

We can also perform PCA for dimensionality reduction to pick the optimal number of feature but than can be done with prior consultation with the product and marketing team.