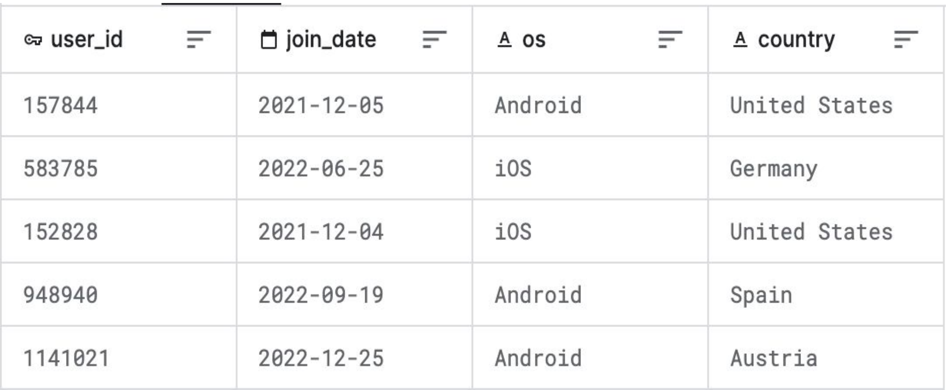
# Final project: Churn Prediction in Mobile Games

## Final Presentation

SAPIENZA
UNIVERSITÀ DI ROMA

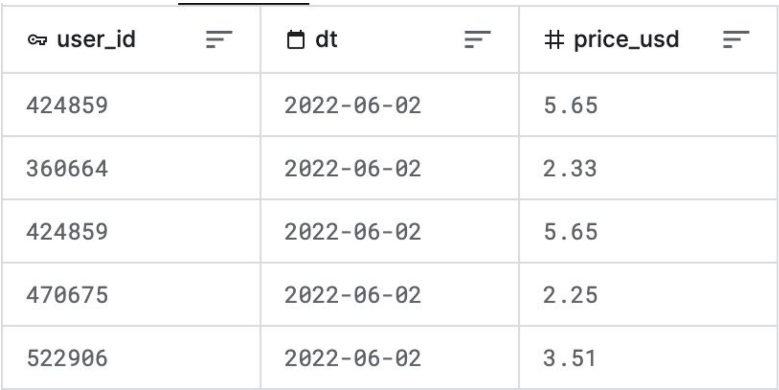
Dila Aslan - 2113310
Selin Erol - 2053197
Umut Altun - 2101934

# Data Sources



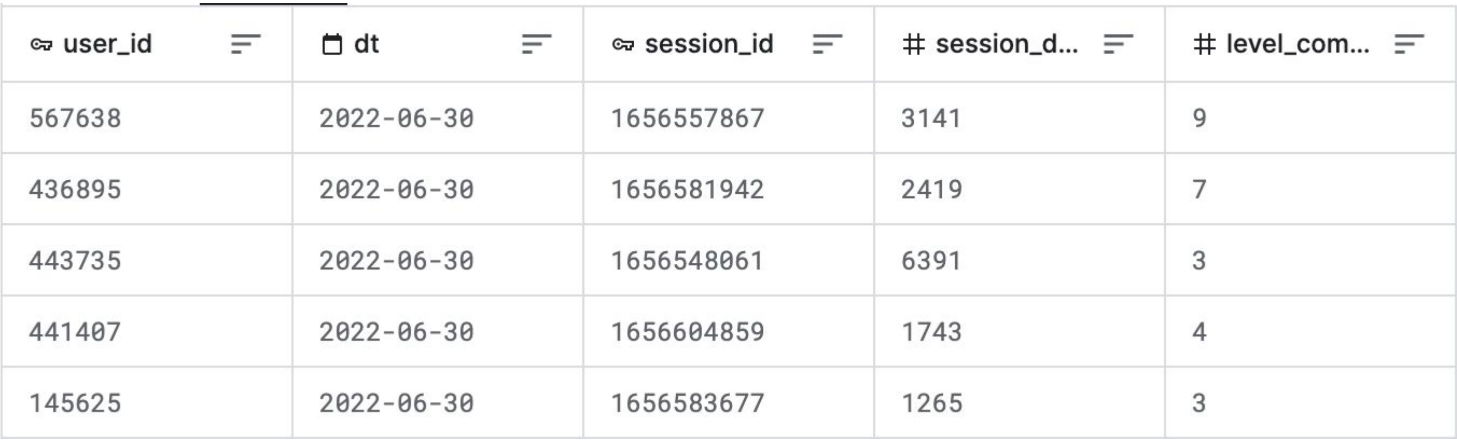
| user_id | join_date | os | country |
|---|---|---|---|
| 157844 | 2021-12-05 | Android | United States |
| 583785 | 2022-06-25 | iOS | Germany |
| 152828 | 2021-12-04 | iOS | United States |
| 948940 | 2022-09-19 | Android | Spain |
| 1141021 | 2022-12-25 | Android | Austria |

User Table, Shape: (6584, 4)



| user_id | dt | price_usd |
|---|---|---|
| 424859 | 2022-06-02 | 5.65 |
| 360664 | 2022-06-02 | 2.33 |
| 424859 | 2022-06-02 | 5.65 |
| 470675 | 2022-06-02 | 2.25 |
| 522906 | 2022-06-02 | 3.51 |

Purchase Table, Shape: (236270, 3)

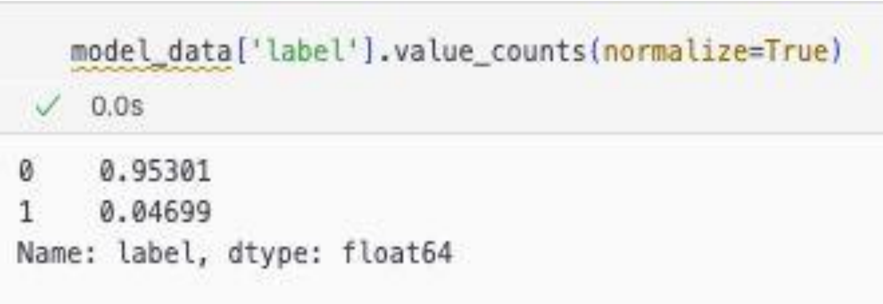


| user_id | dt | session_id | session_d... | level_com... |
|---|---|---|---|---|
| 567638 | 2022-06-30 | 1656557867 | 3141 | 9 |
| 436895 | 2022-06-30 | 1656581942 | 2419 | 7 |
| 443735 | 2022-06-30 | 1656548061 | 6391 | 3 |
| 441407 | 2022-06-30 | 1656604859 | 1743 | 4 |
| 145625 | 2022-06-30 | 1656583677 | 1265 | 3 |

Session Table, Shape: (1699352, 5)

# Labeling Target Column

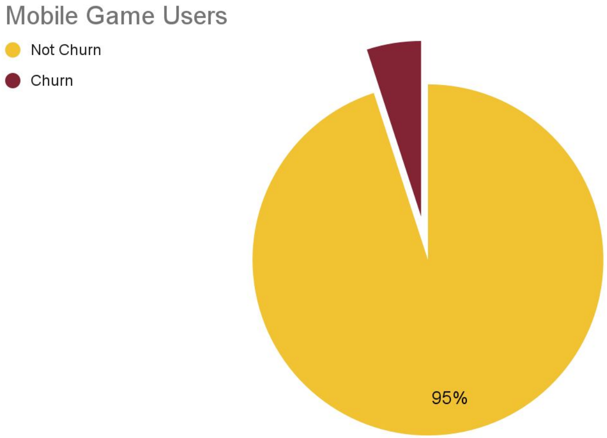
Churn Definition: Churn is defined as the inactivity of a user for more than 3 consecutive days and target variable is defined accordingly.

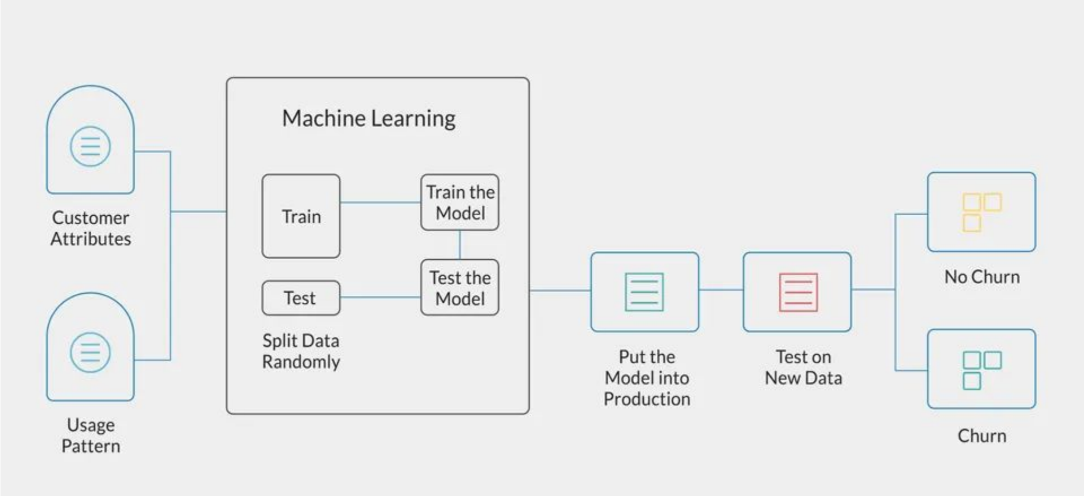


```
model_data['label'].value_counts(normalize=True)
```

✓ 0.0s

```
0    0.95301
1    0.04699
Name: label, dtype: float64
```





# Feature Engineering

## Some Considerations

**User Engagement:**

- Session counts of a user over a period
- Purchases of a user over a period

**Country-Specific Features:**

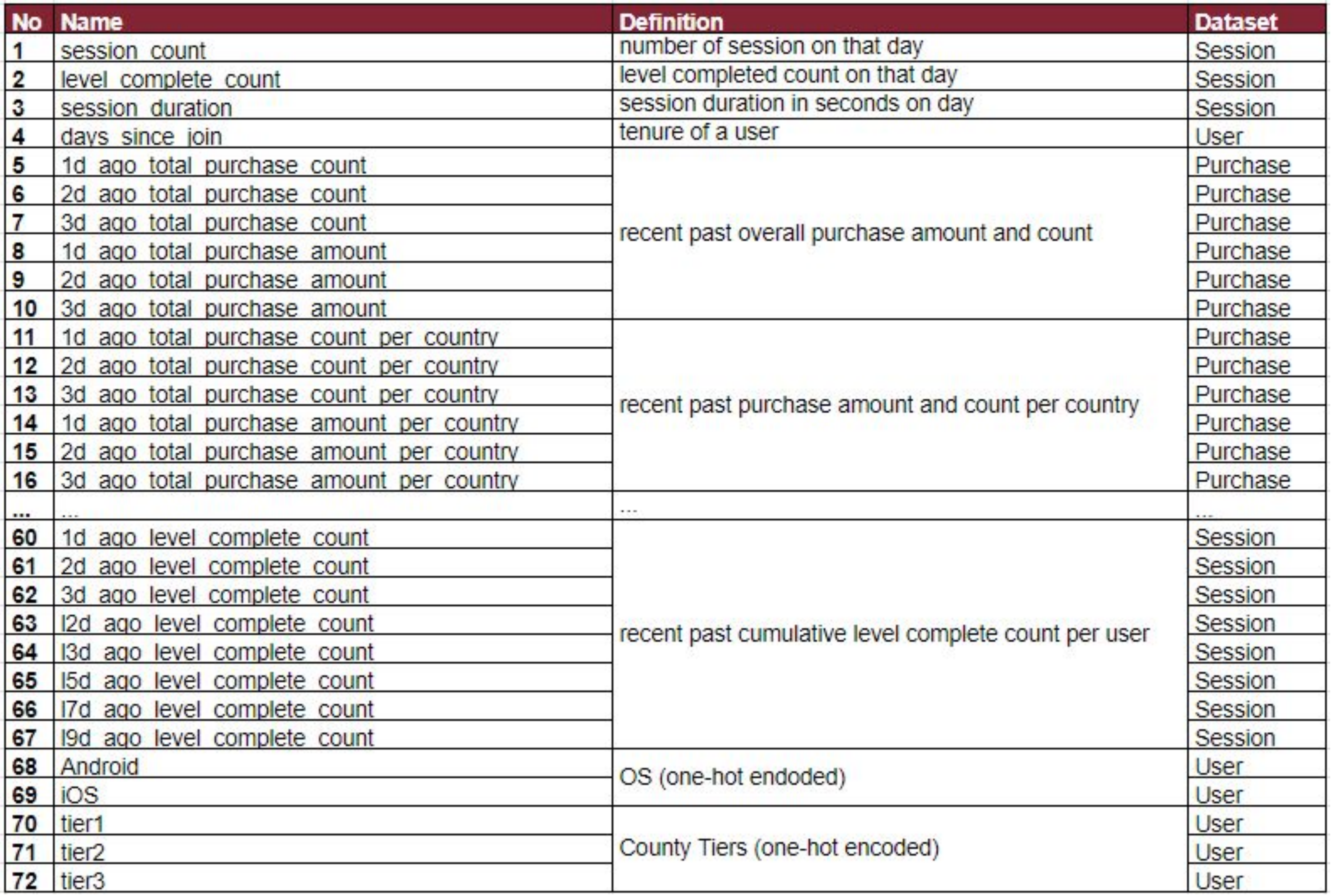
- Session counts in a country over a period
- Purchases of a user in a country over a period

**Marketing Campaigns Launched on the OS Related Features:**

- Session counts in an OS over period
- Purchases in an OS over a period

Features derived from single-day data as well as cumulative-day data have been combined, providing a comprehensive view of user behavior.

# At the end of feature engineering…

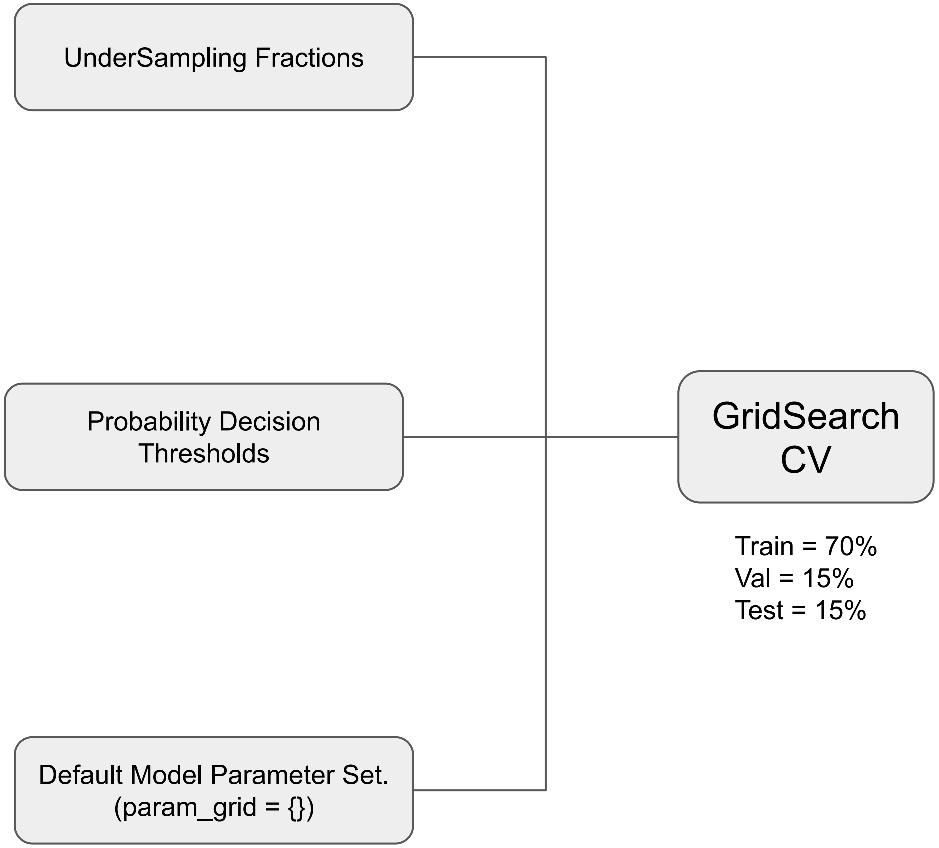


| No | Name | Definition | Dataset |
|---|---|---|---|
| 1 | session count | number of session on that day | Session |
| 2 | level complete count | level completed count on that day | Session |
| 3 | session duration | session duration in seconds on day | Session |
| 4 | days since join | tenure of a user | User |
| 5 | 1d ago total purchase count | recent past overall purchase amount and count | Purchase |
| 6 | 2d ago total purchase count | | Purchase |
| 7 | 3d ago total purchase count | | Purchase |
| 8 | 1d ago total purchase amount | | Purchase |
| 9 | 2d ago total purchase amount | | Purchase |
| 10 | 3d ago total purchase amount | | Purchase |
| 11 | 1d ago total purchase count per country | recent past purchase amount and count per country | Purchase |
| 12 | 2d ago total purchase count per country | | Purchase |
| 13 | 3d ago total purchase count per country | | Purchase |
| 14 | 1d ago total purchase amount per country | | Purchase |
| 15 | 2d ago total purchase amount per country | | Purchase |
| 16 | 3d ago total purchase amount per country | | Purchase |
| ... | ... | ... | ... |
| 60 | 1d ago level complete count | recent past cumulative level complete count per user | Session |
| 61 | 2d ago level complete count | | Session |
| 62 | 3d ago level complete count | | Session |
| 63 | l2d ago level complete count | | Session |
| 64 | l3d ago level complete count | | Session |
| 65 | l5d ago level complete count | | Session |
| 66 | l7d ago level complete count | | Session |
| 67 | l9d ago level complete count | | Session |
| 68 | Android | OS (one-hot endoded) | User |
| 69 | iOS | | User |
| 70 | tier1 | County Tiers (one-hot encoded) | User |
| 71 | tier2 | | User |
| 72 | tier3 | | User |

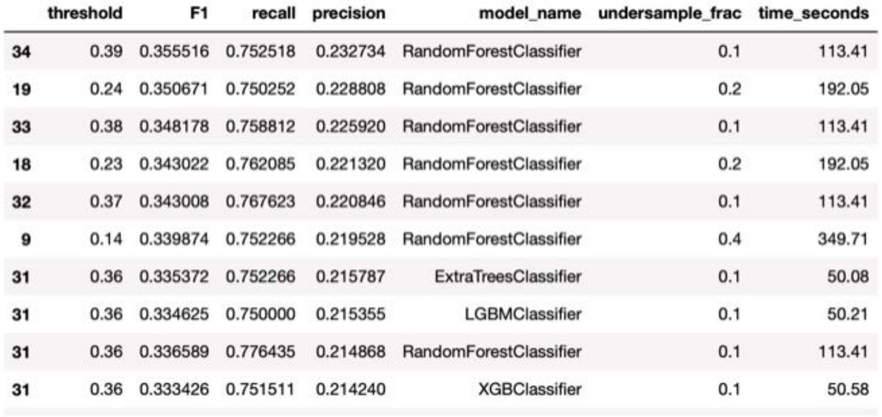
# Modelling

## Base Learner Selection

**Models**
Logistic Reg.
ExtraTrees
LGBM
XGB
Random Forest
AdaBoost


UnderSampling Fractions
Probability Decision Thresholds
Default Model Parameter Set. (param_grid = {})
GridSearch CV

Train = 70%
Val = 15%
Test = 15%



| | threshold | F1 | recall | precision | model_name | undersample_frac | time_seconds |
|---|---|---|---|---|---|---|---|
| **34** | 0.39 | 0.355516 | 0.752518 | 0.232734 | RandomForestClassifier | 0.1 | 113.41 |
| **19** | 0.24 | 0.350671 | 0.750252 | 0.228808 | RandomForestClassifier | 0.2 | 192.05 |
| **33** | 0.38 | 0.348178 | 0.758812 | 0.225920 | RandomForestClassifier | 0.1 | 113.41 |
| **18** | 0.23 | 0.343022 | 0.762085 | 0.221320 | RandomForestClassifier | 0.2 | 192.05 |
| **32** | 0.37 | 0.343008 | 0.767623 | 0.220846 | RandomForestClassifier | 0.1 | 113.41 |
| **9** | 0.14 | 0.339874 | 0.752266 | 0.219528 | RandomForestClassifier | 0.4 | 349.71 |
| **31** | 0.36 | 0.335372 | 0.752266 | 0.215787 | ExtraTreesClassifier | 0.1 | 50.08 |
| **31** | 0.36 | 0.334625 | 0.750000 | 0.215355 | LGBMClassifier | 0.1 | 50.21 |
| **31** | 0.36 | 0.336589 | 0.776435 | 0.214868 | RandomForestClassifier | 0.1 | 113.41 |
| **31** | 0.36 | 0.333426 | 0.751511 | 0.214240 | XGBClassifier | 0.1 | 50.58 |

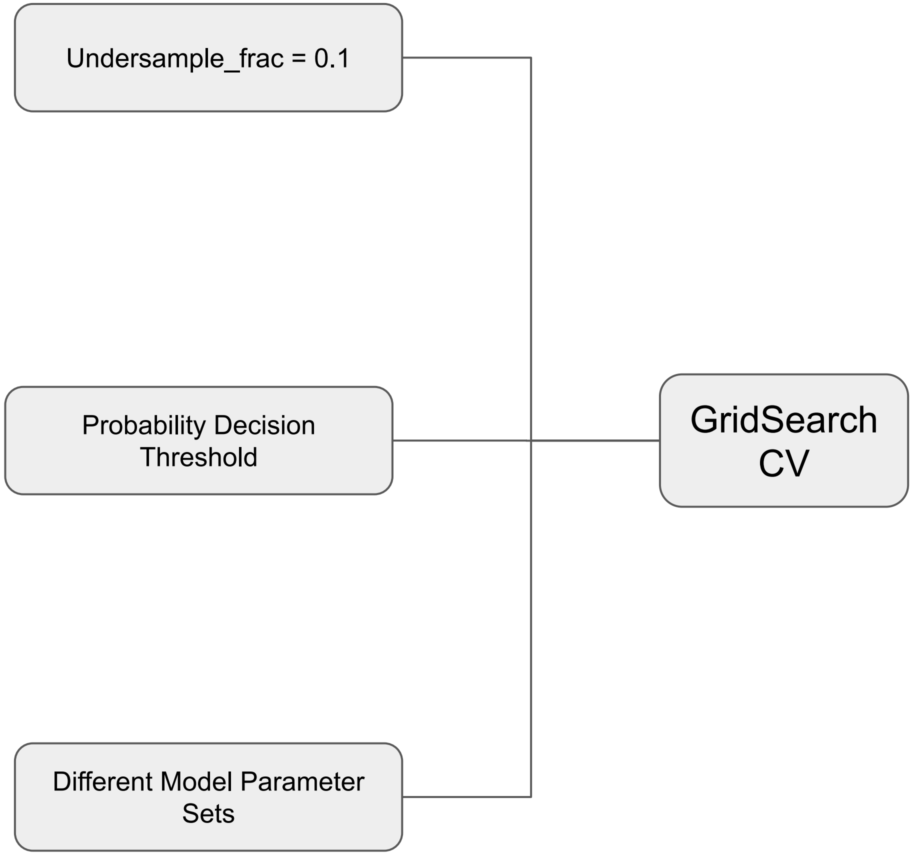
- For business needs, it is better to have a low precision, high recall. So our goal is getting recall minimum 75%. (To cover at least 75% of True Churned Users)
- With this evaluation threshold, best results came with undersample_frac =0.1 for each model type.

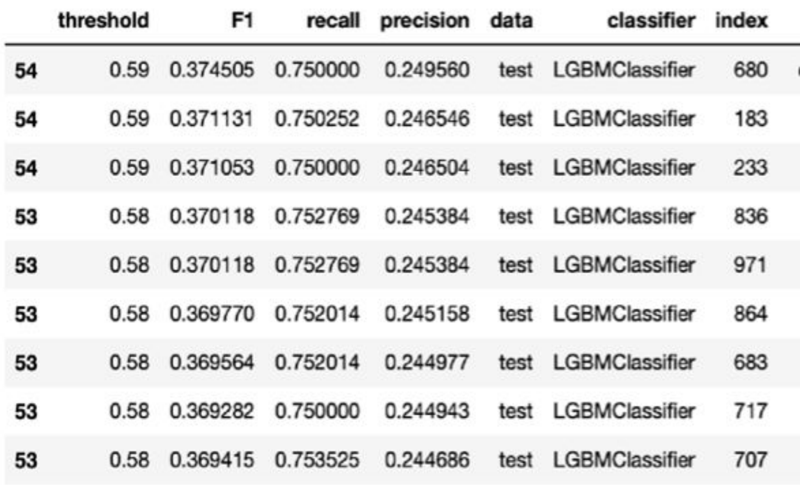
# Modelling

## Best Learner Selection

**Models**

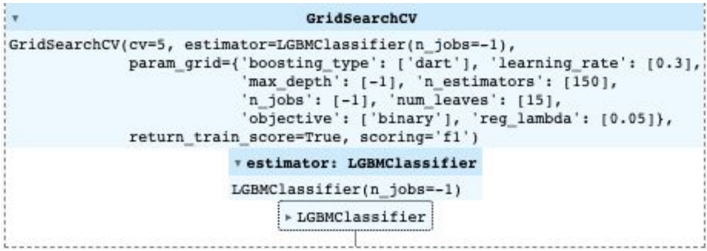
Logistic Reg.
ExtraTrees
LGBM
XGB
Random Forest
AdaBoost


Undersample_frac = 0.1
Probability Decision Threshold
Different Model Parameter Sets
GridSearch CV



| | threshold | F1 | recall | precision | data | classifier | index |
|---|---|---|---|---|---|---|---|
| 54 | 0.59 | 0.374505 | 0.750000 | 0.249560 | test | LGBMClassifier | 680 |
| 54 | 0.59 | 0.371131 | 0.750252 | 0.246546 | test | LGBMClassifier | 183 |
| 54 | 0.59 | 0.371053 | 0.750000 | 0.246504 | test | LGBMClassifier | 233 |
| 53 | 0.58 | 0.370118 | 0.752769 | 0.245384 | test | LGBMClassifier | 836 |
| 53 | 0.58 | 0.370118 | 0.752769 | 0.245384 | test | LGBMClassifier | 971 |
| 53 | 0.58 | 0.369770 | 0.752014 | 0.245158 | test | LGBMClassifier | 864 |
| 53 | 0.58 | 0.369564 | 0.752014 | 0.244977 | test | LGBMClassifier | 683 |
| 53 | 0.58 | 0.369282 | 0.750000 | 0.244943 | test | LGBMClassifier | 717 |
| 53 | 0.58 | 0.369415 | 0.753525 | 0.244686 | test | LGBMClassifier | 707 |

- We get the best F1 Score ( which at least 75% Recall score) with LGBM Classifier.
- Also from time complexity perspective, LGBM surpassed other models.
- Best LGBM Model Parameter Set
  - Undersampling_frac = 0.1, Prob .Threshold = 0.59

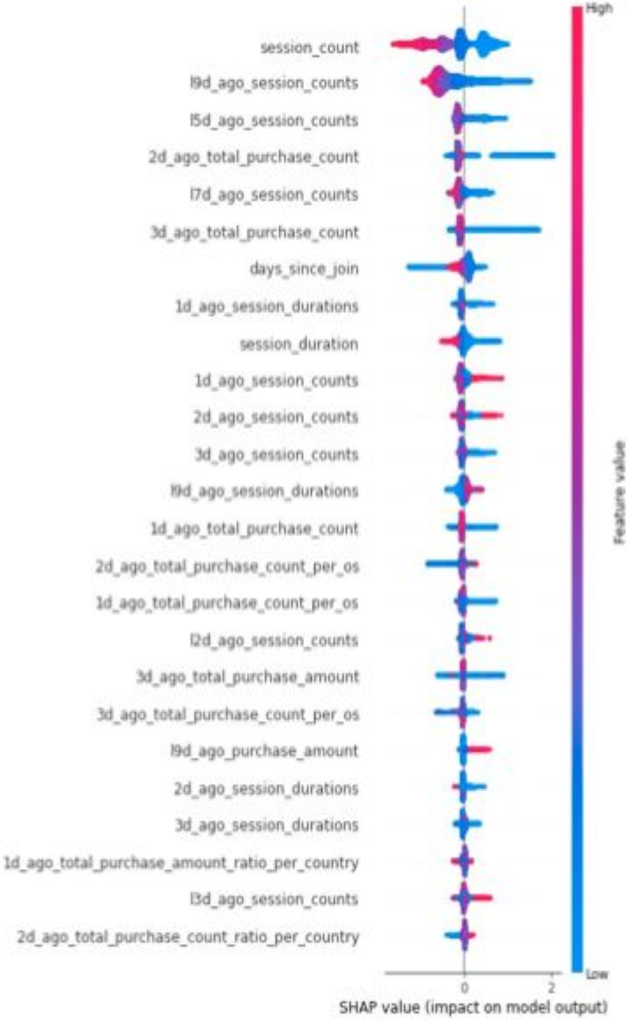


```
▾ GridSearchCV
GridSearchCV(cv=5, estimator=LGBMClassifier(n_jobs=-1),
             param_grid={'boosting_type': ['dart'], 'learning_rate': [0.3],
                         'max_depth': [-1], 'n_estimators': [150],
                         'n_jobs': [-1], 'num_leaves': [15],
                         'objective': ['binary'], 'reg_lambda': [0.05]},
             return_train_score=True, scoring='f1')
                    ▾ estimator: LGBMClassifier
                    LGBMClassifier(n_jobs=-1)
                         ▸ LGBMClassifier
```

# Modelling

## Feature Importance

- Session Count ⬆ Churn Probability ⬇
  - Make users complete as many level as possible.
- Session Duration ⬆ Churn Probability ⬇
  - Avoid Too Hard / Too Easy Levels.
- Purchase Count ⬆ Churn Probability ⬇
  - Make promotions for In App Purchases
- Purchase Amount ⬆ Churn Probability ⬆
  - Decide sensible prices for In App Purchases.


session_count
19d_ago_session_counts
15d_ago_session_counts
2d_ago_total_purchase_count
17d_ago_session_counts
3d_ago_total_purchase_count
days_since_join
1d_ago_session_durations
session_duration
1d_ago_session_counts
2d_ago_session_counts
3d_ago_session_counts
19d_ago_session_durations
1d_ago_total_purchase_count
2d_ago_total_purchase_count_per_os
1d_ago_total_purchase_count_per_os
12d_ago_session_counts
3d_ago_total_purchase_amount
3d_ago_total_purchase_count_per_os
19d_ago_purchase_amount
2d_ago_session_durations
3d_ago_session_durations
1d_ago_total_purchase_amount_ratio_per_country
13d_ago_session_counts
2d_ago_total_purchase_count_ratio_per_country
High
Feature value
Low
SHAP value (impact on model output)