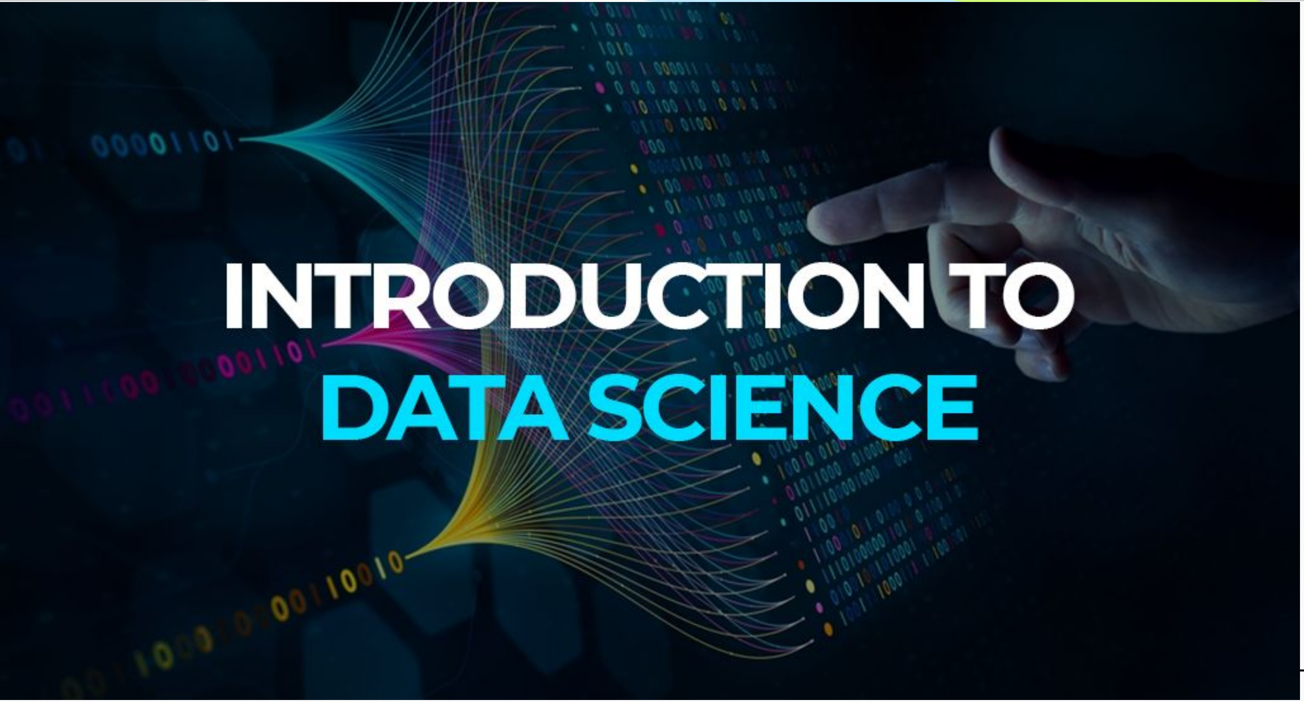
Skoltech

Skolkovo Institute of Science and Technology



Arofenitra Rarivonjy
Gregoire Ouerdane
Team : LES DEUX GENIES

2025 October

The Team : Les deux genies

Presentation, Data Visualization, Feature engineering, modelling

Arofenitra Rarivonjy

Feature engineering, Modelling, Evaluation, RF, XGBoost, GBM

Gregoire Ouerdane

Skoltech

Skolkovo Institute of Science and Technology

Let’s get started!

Project Presentation

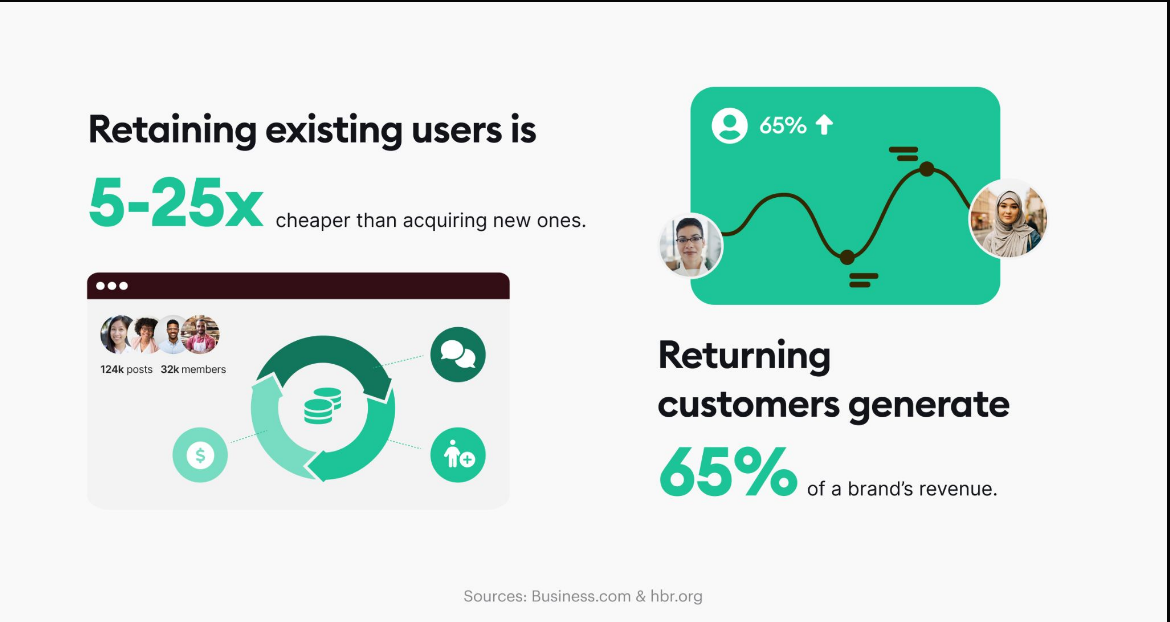
Skolkovo Institute of Science and Technology

Skoltech

PROBLEM DESCRIPTION

Prediction of user churn (inactivity) for a knowledge-sharing platform (Zindi). We need to identify which users are at risk of not engaging with the platform in the coming month.

Importance of the research



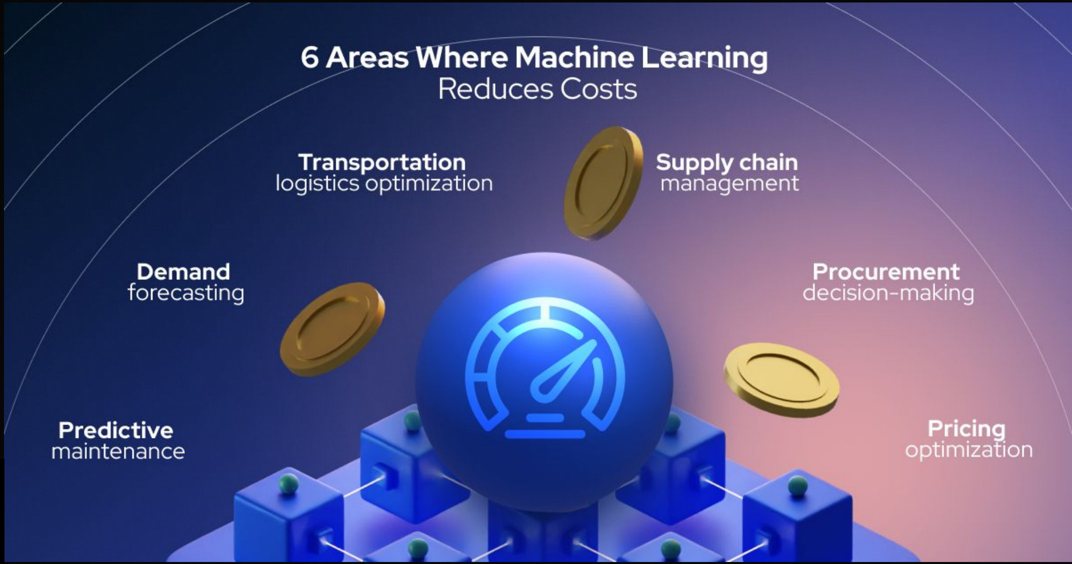
Proactive steps to re-engage non-active users via personalized emails, notifications about new relevant competitions, etc.

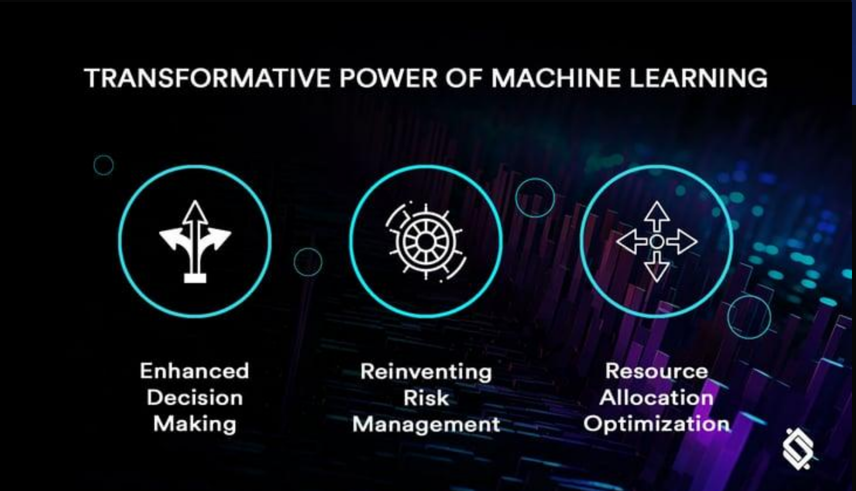
Customer Name	Status	Prediction	Action
Christa Sanders	Active	Active	✓ No Action
Lucas Garcia	Active	Churned	⚠ At-Risk!
Holly Tucker	Inactive	Churned	✓ No Action
Rebecca Presland	Active	Active	✓ No Action
James Moore	Inactive	Active	Marketing Target
Michael Farren	Active	Active	✓ No Action

Skoltech
Skolkovo Institute of Science and Technology

How Machine learning help ?

Reduces cost by focusing on most important

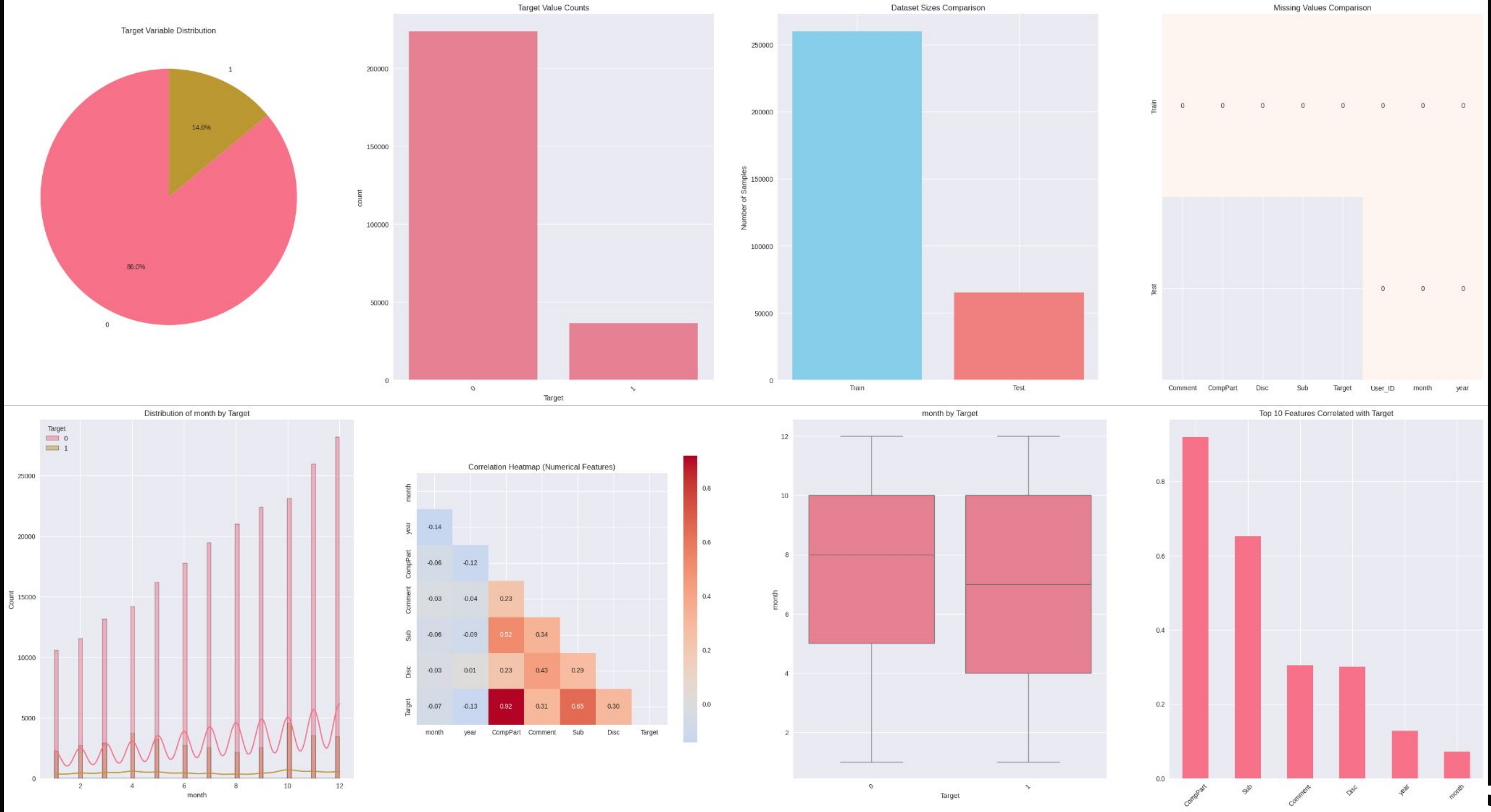




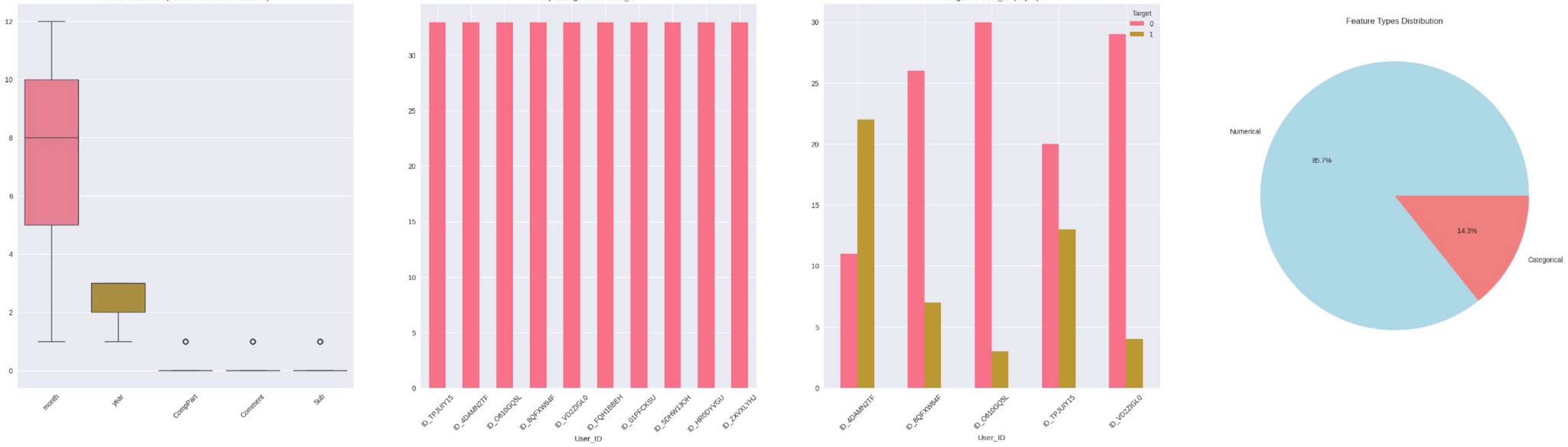
Enhancing decision

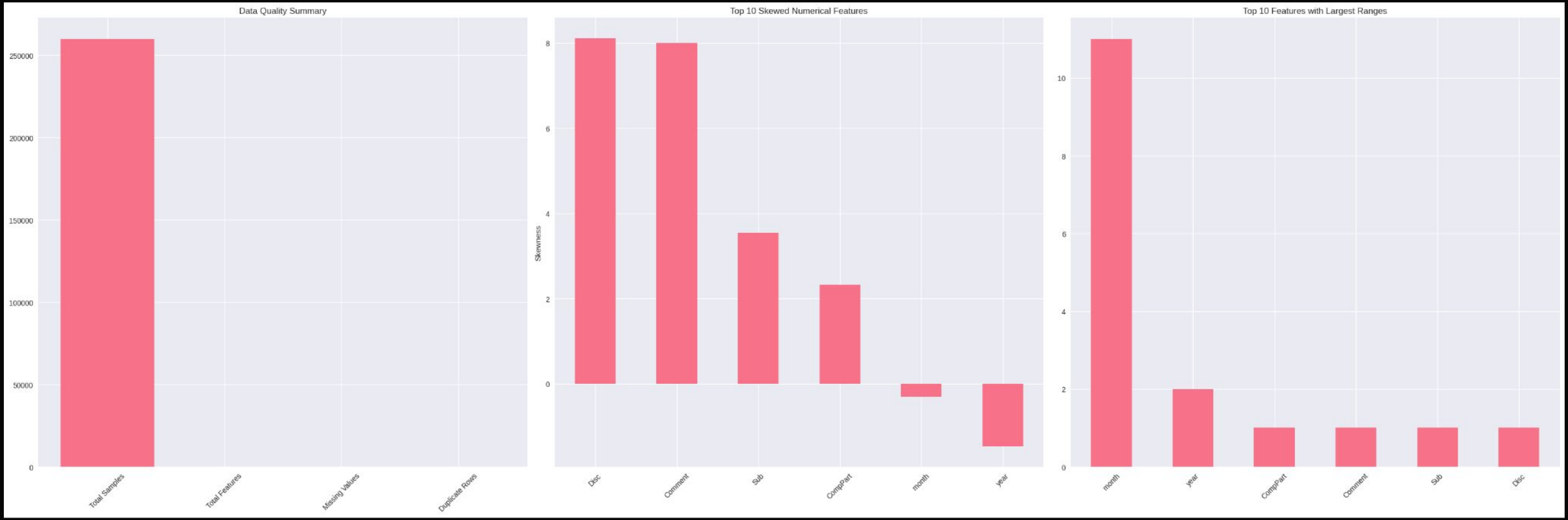
Skoltech
Skolkovo Institute of Science and Technology

Data Visualization

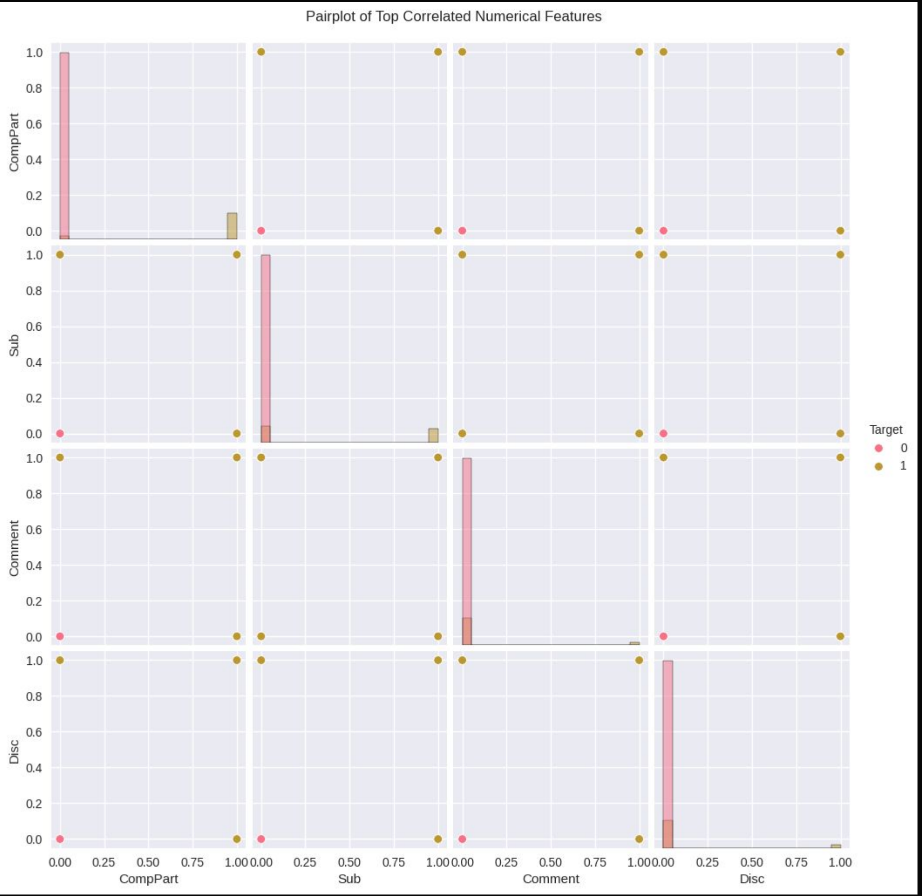


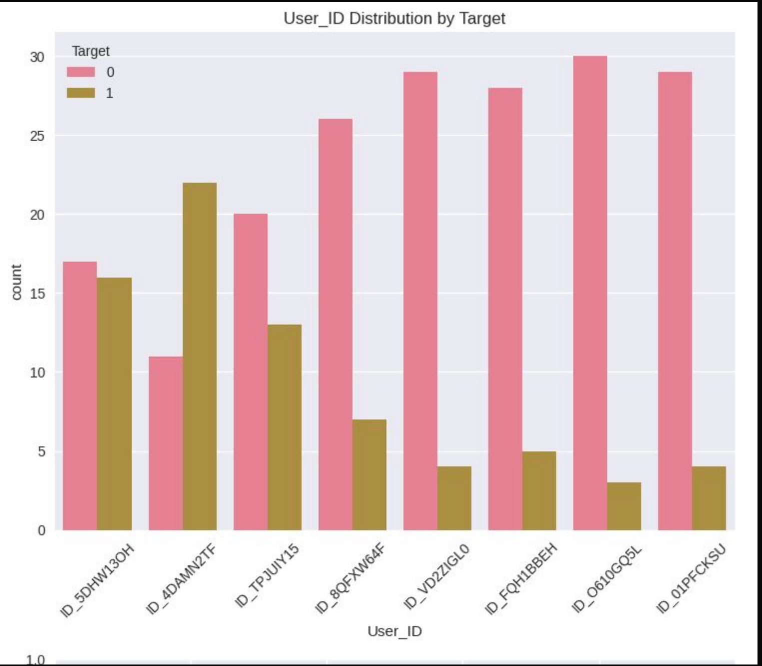
Skoltech
Skolkovo Institute of Science and Technology


Feature Types Distribution
Numerical
85.7%
14.3%
Categorical
Target
0
1
User_ID
User_ID
month
year
CompPart
Comment
Sub


Data Quality Summary
Total Samples
Total Features
Missing Values
Duplicate Rows
Top 10 Skewed Numerical Features
Skewness
Disc
Comment
Sub
CompPart
month
year
Top 10 Features with Largest Ranges
month
year
CompPart
Comment
Sub
Disc

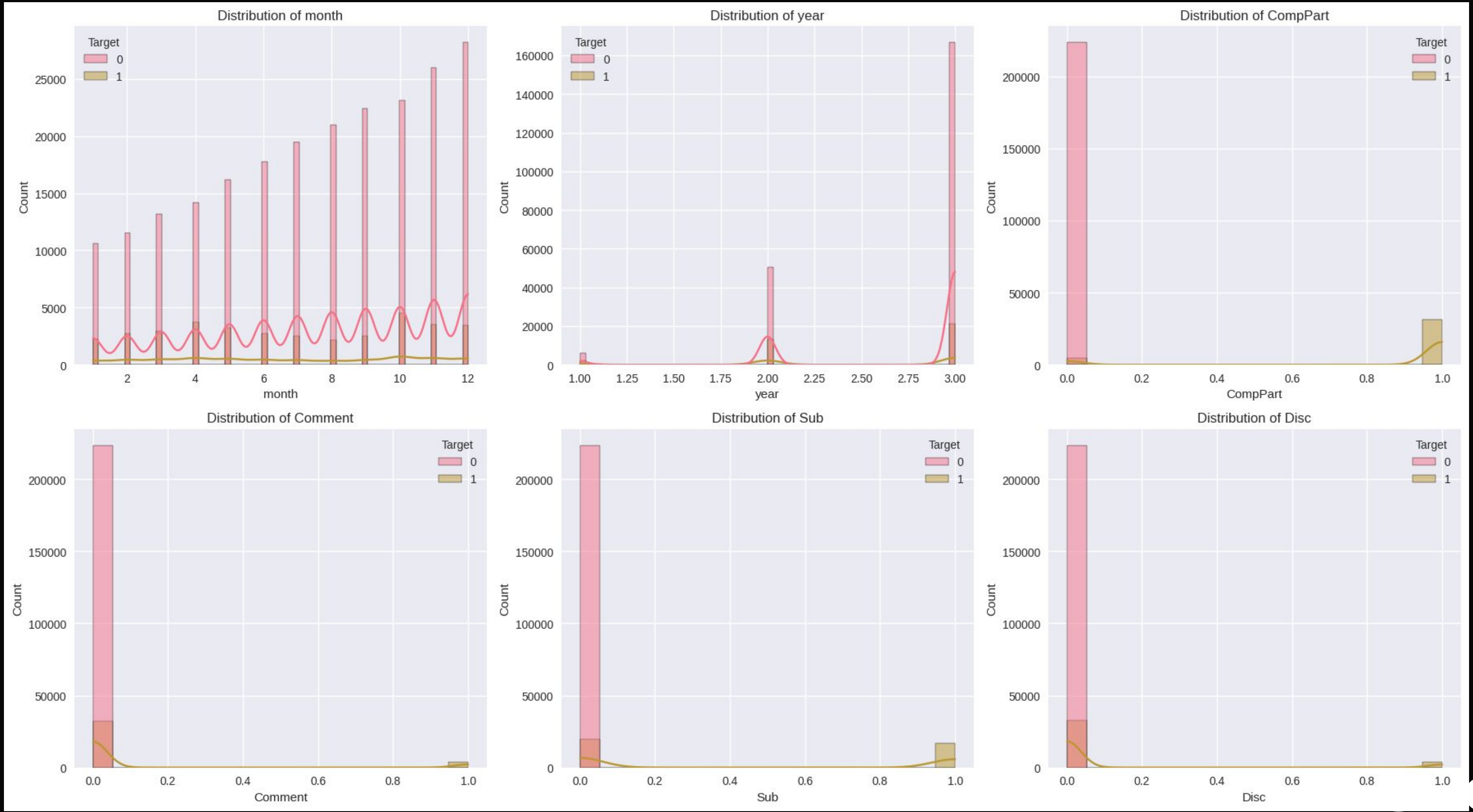
Data Visualization





Skoltech
Skolkovo Institute of Science and Technology

Data Visualization



Skoltech
Skolkovo Institute of Science and Technology

Data Summary and Methods

- Severe Class Imbalance Class 0 (Inactive): 86.0% (223,526 samples) ; Class 1 (Active): 14.0% (36,306 samples)
- Accuracy would be a terrible metric (a model that always predicts "0" would get 86% accuracy but be useless). We will use F1-score or ROC-AUC. Need for handling class imbalance by SMOTE.
- Feature Types Categorical: User_ID (object type)
- Numerical: month, year, CompPart, Comment, Sub, Disc
- Dropping User_ID: adding it may cause overfitting

Findings

- Strong Predictors CompPart (Competition Participation): 0.919 correlation with Target → EXTREMELY STRONG PREDICTOR
- Sub (Submissions): 0.654 correlation with Target → STRONG PREDICTOR
- Comment and Disc (Discussion): ~0.30 correlation → MODERATE PREDICTORS
  - Weak/Negligible Predictors: month and year: Very weak correlations (-0.07 to -0.13)

🔍 Feature Relationships: No highly correlated feature pairs (all < 0.7), so no redundancy issues

- CompPart and Sub have inter-feature correlation (0.520) - submitting requires participation

📈 Activity Feature Distribution: All activity features are binary (0/1) and highly imbalanced:

- Comment and Disc: Very rare (only 1.5% of users)
- CompPart: 12.1% of users participate
- Sub: 6.5% of users make submissions

First implementation

F1 Score: 1.0

Classification Report:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	44706
1	1.00	1.00	1.00	7261
accuracy			1.00	51967
macro avg	1.00	1.00	1.00	51967
weighted avg	1.00	1.00	1.00	51967

- Model : Random Forest (balanced class weight)
- Existence of data leakage
- The provided features in train.csv directly reveal the target variable”

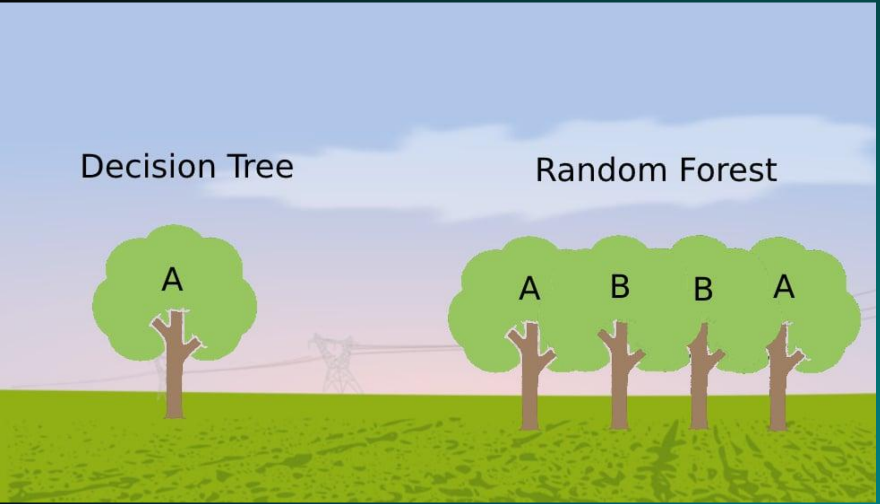
Investigating data leakage

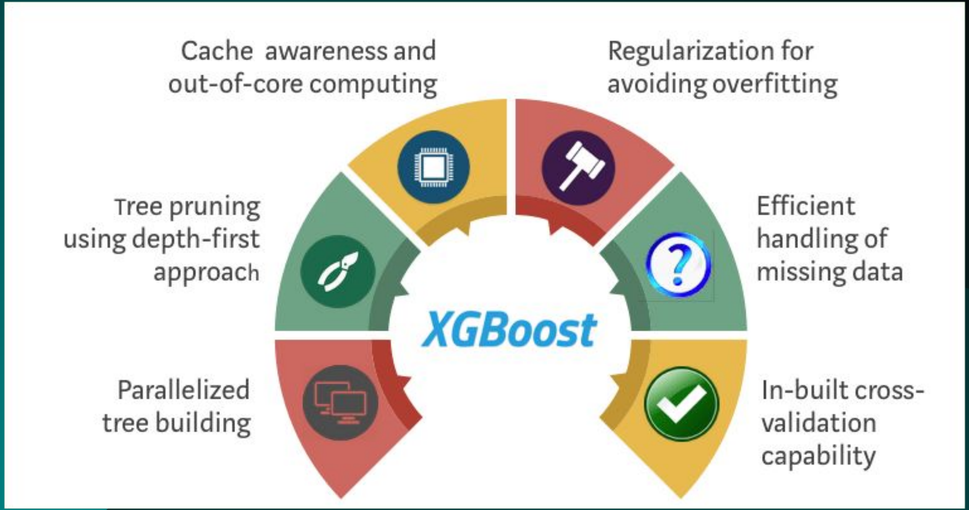
Target	CompPart		Comment		Sub		Disc	
	0	1	0	1	0	1	0	1
0	178820	0	178820	0	178820	0	178820	0
1	3959	25086	25937	3108	15556	13489	26023	3022

Solution : Rebuild from scratch raw data to build meaningful features that do not leak.

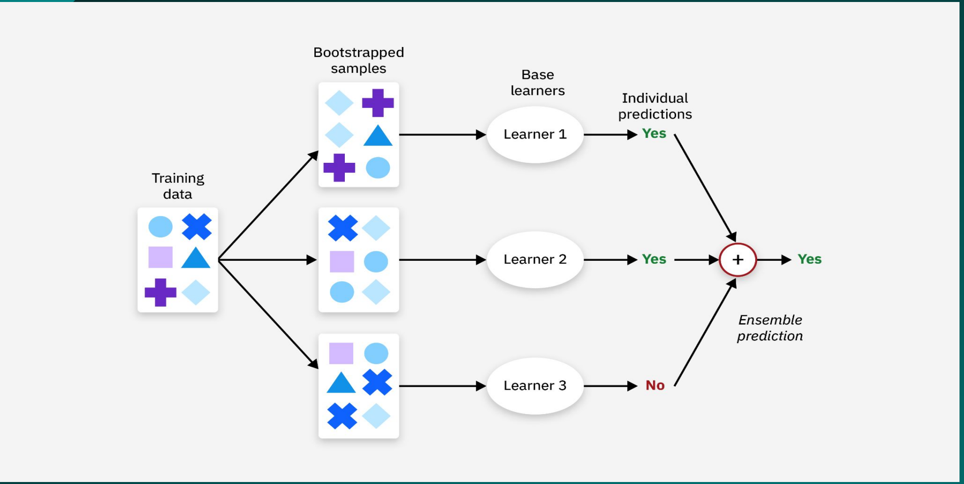
users: (22407, 8)
competitions: (154, 17)
comp_participation: (48565, 7)
submissions: (375763, 6)
discussions: (6211, 6)
comments: (11751, 4)

- Static features we'll use: UserID ; Country ; Points ; user_tenure_months ; FeatureX ; FeatureY
- Merging features with targets...
- Final modeling dataset shape: (20218, 7)
- Class distribution in final dataset: Target (1 <-> 0.751; 0 <-> 0.249)


Decision Tree
Random Forest
A
A
B
B
A


Cache awareness and
out-of-core computing
Regularization for
avoiding overfitting
Tree pruning
using depth-first
approach
Efficient
handling of
missing data
XGBoost
Parallelized
tree building
In-built cross-
validation
capability


Gradient Boosting Algorithm
www.educba.com


Training
data
Bootstrapped
samples
Base
learners
Individual
predictions
Learner 1
Yes
Learner 2
Yes
Learner 3
No
Yes
Ensemble
prediction

Metrics

- Categorical columns: ['UserID', 'Country', 'Points']
- Numerical columns: ['user_tenure_months', 'FeatureX', 'FeatureY']
- Features shape: (20218, 6)
- Target shape: (20218,)
- Train set: (16174, 6)
- Validation set: (4044, 6)
- Train class distribution: {1: 0.751, 0: 0.249}
- Val class distribution: {1: 0.751, 0: 0.249}

Random Forest Output

F1 Score: 0.7081
ROC AUC: 0.7722

	precision	recall	f1-score	support
0	0.40	0.86	0.55	1005
1	0.93	0.57	0.71	3039
accuracy			0.64	4044
macro avg	0.66	0.72	0.63	4044
weighted avg	0.80	0.64	0.67	4044

Top 10 most important features:

	feature	importance
2	FeatureY	0.104580
0	user_tenure_months	0.055111
16277	Country_ID_Q02	0.050819
16313	Points_group 3	0.050304
16276	Country_ID_PLTE	0.049867
16197	Country_ID_5OWN	0.029669
16245	Country_ID_HIXK	0.020542
16246	Country_ID_HWRH	0.017824
1	FeatureX	0.016830
16311	Country_ID_ZXLV	0.014823

Optimized Model

1. Creating advanced Features

Original features: ['Country', 'Points', 'user_tenure_months', 'FeatureX', 'FeatureY', 'Target']

New features created: ['tenure_group', 'points_group', 'log_tenure', 'sqrt_points', 'points_per_month', 'high_points_long_tenure', 'new_user_high_points', 'tenure_squared', 'points_squared', 'feature_ratio', 'activity_score', 'is_new_user', 'is_established_user']

Total features after engineering: 18

2. Advanced Preprocessing and selection of most important features

- Final feature matrix shape: (20218, 18)
- Training set: (16174, 18), Validation set: (4044, 18)
- Selected 7 most important features: 'Country', 'Points', 'points_group', 'log_tenure', 'sqrt_points', 'points_per_month', 'points_squared']

3. Hyperparameters tuning and building of super ensemble

1. **Tuning Random Forest model ...**

Fitting 3 folds for each of 12 candidates, totalling 36 fits
Best rf params: {'class_weight': 'balanced', 'max_depth': None, 'min_samples_split': 2, 'n_estimators': 200}
Best rf CV F1: 0.7988

2. **Tuning Gradient Boosting ...**

Fitting 3 folds for each of 8 candidates, totalling 24 fits
Best gb params: {'learning_rate': 0.1, 'max_depth': 4, 'n_estimators': 200}
Best gb CV F1: 0.8610

3. **Tuning XGBoost ...**

Fitting 3 folds for each of 16 candidates, totalling 48 fits
Best xgb params: {'learning_rate': 0.1, 'max_depth': 4, 'n_estimators': 200, 'subsample': 0.9}
Best xgb CV F1: 0.8630

RANDOM FOREST			Confusion matrix	
F1-score	**roc-auc**	**Acc.**	842	163
0.7880	0.8389	0.723	957	2082
	Precision	**Recall**	**F1-**	**Support**
0	0.47	0.84	0.60	1005
1	0.93	0.69	0.79	3039
accuracy			0.72	4044
Macro avg	0.70	0.76	0.69	4044
Weighted avg	0.81	0.72	0.74	4044

Gradient Boosting			Confusion matrix	
F1-	**roc-auc**	**Acc.**	474	531
0.859	0.847	0.781	353	2686
	Precision	**Recall**	**F1-**	**Support**
0	0.57	0.47	0.52	1005
1	0.83	0.88	0.86	3039
acc			0.78	4044
Macro avg	0.70	0.68	0.69	4044
Weighted avg	0.77	0.78	0.77	4044

XGBOOST			Confusion matrix	
F1-	**roc-auc**	**Acc.**	497	508
0.86	0.847	0.785	362	2677

	Precision	**Recall**	**F1-**	**Support**
0	0.58	0.49	0.53	1005
1	0.84	0.88	0.86	3039
acc			0.78	4044
Macro avg	0.71	0.69	0.70	4044
Weighted avg	0.78	0.78	0.78	4044

ENSEMBLE RESULTS		
F1-	**roc-auc**	**Acc.**
0.8464	0.8468	0.7762

FEATURE IMPORTANCE ANALYSIS AND SUMMARY

TOP 15 MOST IMPORTANT FEATURES (XGBoost):

	feature	importance
1	Points	0.962874
3	log_tenure	0.022328
0	Country	0.011952
5	points_per_month	0.002846
2	points_group	0.00
4	sqrt_points	0.00
6	points_squared	0.00

Baseline F1 Score: 0.7705
Best Optimized F1 Score: 0.8602
Improvement: +0.0897 (11.6%)

🎯 BEST OVERALL MODEL: XGBoost
🎯 BEST F1 SCORE: 0.8602

STACKING ENSEMBLE RESULTS:
F1 Score: 0.8523
ROC AUC: 0.8311

🏆 FINAL BEST F1 SCORE: 0.8602
🏆 TOTAL IMPROVEMENT: +0.0897 (11.6%)

CONCLUSION

1. DATA LEAKAGE DISCOVERY:
   - Initial model had perfect F1-score (1.0) due to data leakage
   - Activity features directly revealed the target variable
   - This taught us the importance of proper temporal feature engineering

2. VALID SOLUTION:
   - Switched to static user features only (no leakage possible)
   - Used: Country, Points, User Tenure, FeatureX, FeatureY
   - Achieved realistic basic F1-score of 0.7705 and optimized F1-score of 0.8602

3. BUSINESS IMPACT:
   - Model can identify users likely to become inactive based on profile characteristics
   - Platform can target retention efforts more effectively
   - Countries like ID_Q02 and longer tenure users show different activity patterns

4. TECHNICAL ACHIEVEMENT:
   - Built a valid predictive model without data leakage
   - Handled class imbalance using class weights
   - Used appropriate metrics (F1, ROC-AUC) for imbalanced data