

SPEC2MODEL Challenge — Submission Sheet

GDGOC Silicon University — ZYGON x Neosis Annual Fest

Section 1 — Team Information

- Team Name: Team STATS
- Problem Chosen: B
- Member 1: Arpit Kumar Nayak, 22BECG33, ECE
- Member 2: Jitesh Bhakat, 22BCSI23, CSE
- Contact (any one member): 7847050186

Section 2 — Data Strategy

Did you use any external data beyond the provided train.csv?

- No

If yes, fill the table below for each source:

Source Name	URL	Rows Used	Why It Was Relevant

How did you verify the external data was clean and trustworthy?

Answer:

Section 3 — Data Cleaning and Preprocessing

How did you handle missing values? Which columns had them and what did you do?

Answer: Used simple imputations on grouped columns on median values.

Did you find any noisy or suspicious labels in the training data? What did you do about them?

Answer: used log transform and robust scaler to handle outliers.

Was the data imbalanced? How did you handle it (if at all)?

Answer: No

Any other transformations you applied (scaling, encoding, outlier removal)?

Answer: Robust Scaling

Section 4 — Feature Engineering

What were your top 3 most important features and why?

Rank	Feature	Why It Matters
1	competitor_app_installed	Highest weight value
2	price_increase_experienced	
3	session_trend_30d	

Did you create any new features that were not in the original dataset?

Answer: No

Did you drop any features? Which ones and why?

Answer: Yes, "id" - not required.

"avg_session_duration_min" - high correlated feature

Did you check for correlations or feature interactions? What did you find?

Answer: Yes "avg_session_duration_min" - high correlated feature (VIF : 6.57)

Section 5 — Model Selection and Training

What is your final model?

Answer: XGBoost

What other models did you try before this? Why did you pick the final one over them?

Model Tried	Validation Score	Why You Rejected / Kept It
SVM	0.51	I got better results than this model
Random Forest	0.53	

What hyperparameters did you tune? What values did you settle on?

Answer:n_estimators=300,

```
max_depth=3,  
learning_rate=0.05,  
subsample=0.8,  
colsample_bytree=0.7,  
gamma=0,  
min_child_weight=1,  
reg_alpha=0,  
reg_lambda=2,  
objective="multi:softprob",  
num_class=4,  
eval_metric="mlogloss",  
random_state=42,  
n_jobs=-1
```

How did you validate your model? (train/test split, cross-validation, etc.)

Answer: K- Fold Cross Validation

What was your best validation score before submitting?

Answer: 0.54 -> F1 Macro

0.57 -> accuracy

Section 6 — Honest Reflection

What did you try that did NOT work?

Answer: Couldn't improve evaluation parameters

What are the known limitations of your model?

Answer: It didn't work best on missing values

What was the hardest part of this challenge for your team?

Answer: Handling missing values and outliers

If you had 6 more hours, what would you do differently?

Answer: Preprocessing Better

Section 7 — Team Collaboration

Member

What They Worked On

Arpit Kumar Nayak

Jitesh Bhakat

Section 8 — External Data Declaration

I confirm that all external data sources used are listed in Section 2 and are publicly available.

- No