Submission Sheet

Fill this out during the event and include it as submission_sheet.pdf (or .md) in your PR.

Section 1 — Team Information

- **Team Name: Silent Predictor**
- **Problem Chosen:** A
- **Member 1:** Ashutosh Nayak,23bcsb36,C.S.E
- **Member 2:** Umesh Kumar Sahoo,23bcsc69,C.S.E
- **Member 3:** Himanshu Kumar,23bcsg57,C.S.E
- **Contact (any one member): 8294225202**

Submission Sheet

Section 2 — Data Strategy

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

- **No**

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

Answer:
We did not use any external data. We relied entirely on the provided training dataset to ensure consistency in distribution between training and test sets.

Section 3 — Data Cleaning and Preprocessing

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

Answer:

- Numerical columns were filled using median imputation to avoid distortion from outliers.
- Categorical columns were filled with the value "missing" to preserve information.
- Boolean columns were converted to integer (0/1) format.

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

Answer:
No obviously incorrect labels were found. Some edge cases (e.g., short distance but marked late) were kept because they likely represent real-world unpredictability.

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

Answer:
Yes, the minority class (late) was approximately 35% of the majority class.
Instead of using SMOTE, we used scale_pos_weight in XGBoost to balance the loss function.
This avoids introducing synthetic noise while improving minority class recall.

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

Answer:

- OneHotEncoding was applied to categorical variables.
- No scaling was required for XGBoost.
- No aggressive outlier removal was performed to preserve real-world extreme delivery cases.

Section 4 — Feature Engineering

What were your top 3 most important features and why?

Rank	Feature	Why It Matters
1	Num_restaurant_avg_prep_min	Combines traffic, weather, and peak hour — represents real delivery pressure
2	Num_traffic_risk_score	Measures how much prep time consumes estimated time
3	Num_distance_km	Captures unrealistic delivery expectations

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

Answer:
Yes. We engineered interaction-based features including:

- Grid_lock_long
- Bad_weather

- High_risk_time
- Traffic_risk_score

These features capture real-world delivery pressure and interactions.

Did you drop any features? Which ones and why?

Answer:
The id column was dropped because it has no predictive value.

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

Answer:
Single features showed overlapping distributions between classes.
Delay is caused by combinations such as heavy rain + gridlock + long distance.
Therefore, interaction features significantly improved performance.

Section 5 — Model Selection and Training

What is your final model?

Answer:
Random Forest Classifier

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**
Logistic Regression	Lower	Could not capture nonlinear interactions
Random Forest	Best	Strong generalization and best Macro F1
XGBoost	Moderate	Slight overfitting observed

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

Answer:
We tuned:

- max_depth
- n_estimators
- min_sample_split

- min_samples_leaf

Final configuration:

- max_depth :
- n_estimators
- min_sample_split
- min_samples_leaf

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

Answer:
We used Stratified 5-fold Cross-Validation to preserve class distribution and evaluated using Macro F1 Score.

What was your best validation score before submitting?

Answer:
[57%]

Section 6 — Honest Reflection

What did you try that did NOT work?

Answer:

- SMOTE oversampling slightly reduced validation performance.
- Deep trees (max_depth > 6) caused overfitting.
- Neural networks performed worse than tree-based models.

What are the known limitations of your model?

Answer:

- Relies heavily on interaction patterns.
- May struggle with unseen extreme operational combinations.
- Does not explicitly model temporal trends beyond hour/day features.

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

Answer:
Balancing minority class recall while maintaining strong overall Macro F1 without introducing overfitting.

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

Answer:

- Perform more structured hyperparameter search.
- Implement soft-voting ensemble of top models.
- Explore SHAP for deeper feature importance analysis.

Section 7 — Team Collaboration

Member	What They Worked On
Member 1	Data cleaning & preprocessing
Member 2	Feature engineering
Member 3	Model tuning & validation

Section 8 — External Data Declaration

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

- **Yes**

SPEC2MODEL Challenge — GDGOC Silicon University — Zygon x Neosis Annual Fest