



Audience - Questions and Answers

Predicting Customer Churn

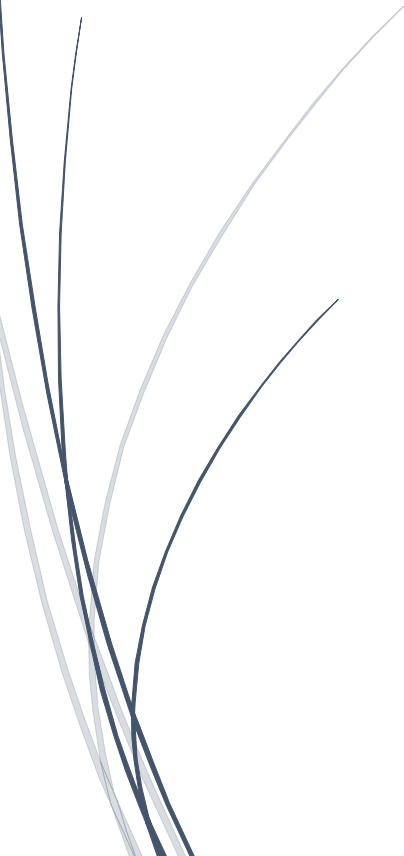
Telecommunications Company

(Project 02 – Milestone 03)

Bellevue University
DSC680 – Applied Data Science

Submitted By:
Debabrata Mishra

Instructor:
Amirfarrokh Iranitalab



Here are ten common questions and answers that an audience might ask regarding the analysis of Water Safety Analyzer:

1. How did you handle the class imbalance between potable and non-potable water samples, and why did you choose ADASYN for oversampling?

To address the class imbalance, we used ADASYN (Adaptive Synthetic Sampling). This method was chosen because it generates synthetic data points for the minority class, improving overall model performance and balance. Unlike SMOTE, ADASYN focuses on creating samples near hard-to-classify instances, enhancing the model's sensitivity to the minority class.

2. What features or parameters had the most significant impact on predicting water potability?

Key features like turbidity, pH, and hardness were crucial for predicting water potability. These parameters were essential because they directly reflect water quality, which is critical for determining whether the water is potable.

3. What challenges did you face during data preprocessing, and how did they impact the model's performance?

Data preprocessing challenges included dealing with missing values and normalizing features. These issues could have introduced bias or inconsistencies in the model. To mitigate these effects, we applied techniques such as imputation for missing data and feature scaling to standardize values.

4. Did you observe any specific geographic or seasonal trends in water potability?

Yes, the analysis showed that water potability varied across different regions and seasons. Certain geographic areas or times of the year had higher levels of contamination, providing valuable insights for targeted water quality management and policymaking.

5. What ethical considerations did you take into account to ensure fairness and transparency in your model's predictions?

We implemented several measures to ensure fairness, including data privacy protections, efforts to reduce bias in the model, and ensuring transparency in predictions. Regular audits and fairness checks were also performed to ensure the model made equitable and unbiased decisions.

6. What are the next steps to further improve the accuracy and applicability of predictive models in water quality management?

The next steps include incorporating additional features, such as real-time environmental data, refining hyperparameters, and using ensemble methods. Continuous model

monitoring and periodic updates will also be crucial for maintaining high accuracy over time.

7. What criteria did you use to select the machine learning algorithms, and why were these chosen over others?

We selected algorithms based on their ability to handle imbalanced data, accuracy, and interpretability. Random Forest and XGBoost were preferred because of their robustness and capability to capture complex relationships within the data.

8. Why did Random Forest outperform models like SVM and KNN in terms of accuracy and precision for predicting water potability?

Random Forest outperformed other models due to its ensemble approach, which reduces overfitting and improves generalization. By aggregating predictions from multiple decision trees, it delivers better accuracy and precision compared to models like SVM and KNN.

9. Given the dynamic nature of water quality and environmental factors, how do you plan to validate and update your predictive models over time?

To handle dynamic changes, we plan to periodically retrain the model with new data, continuously evaluate its performance, and incorporate real-time data updates. This strategy will help keep the model accurate and up to date.

10. What were the key challenges during model deployment, particularly with the Random Forest model, and how did you address them?

Key challenges included scaling the model for production and integrating it with existing systems. These were resolved by optimizing the model for deployment and ensuring smooth integration with data pipelines and monitoring tools.