



## Assistants

THREAD thread\_Hps5ETvRJevYqcTMQExiBYIf

32335 tokens



1

### User

let's get started.

 Resume\_JunJie\_Zhang\_2025.docx

### DS Interviewer Assistant

Great! Let's get started with the interview.

**Question 1:** Can you tell me more about your role at Bayer AG as a Senior Statistical Data Scientist? Specifically, I'm interested in the Advanced Crop Yield Forecasting Model you developed. What was the goal of this model, and what methodologies did you apply?

### User

sure. we need to predict the yields of different hybrids of seeds by site, which give directions of how many acres we should plant each year. we used multiple machine learning algorithms to do that

### DS Interviewer Assistant

That sounds interesting. Can you elaborate on the specific machine learning algorithms you used for the yield forecasting model? Additionally, how did you handle feature selection, and what features did you find to be the most impactful in your predictions?

### User

BLUP, random forest, SVM XGboost and NN. feature selection is critical, by comparing and experimental trial, we select features that are most important. Some genetic and biological features are most important.

### DS Interviewer Assistant

It's great to hear that you employed a diverse range of algorithms.

**Follow-up Question:** Could you explain how you implemented feature selection using experimental trials? What criteria did you use to determine the importance of genetic and biological features, and how did these features impact model performance? Moreover, did you encounter any challenges with multicollinearity or overfitting, and how did you address them?

**User**

we try to drop the features and compare the impact to the model performance. we calculate the feature importance after modeling to evaluate their importance. yes we have a lot of multivolnearity and overfitting issues. what we do are, drop some collinearity features, do feature engineering, tune hyperparameters (ie keep the depth less than 10 for random forest).

**DS Interviewer Assistant**

It sounds like you have a solid approach in managing feature selection and addressing issues like multicollinearity and overfitting.

**Further Inquiry:** Regarding hyperparameter tuning, how did you decide on the specific hyperparameters to tune for each algorithm? Did you use any specific search methods, like grid search or random search, to optimize these parameters? Additionally, how did you validate the model's performance after tuning? Could you also highlight the evaluation metrics you prioritized?

**User**

hyperparameters dependet on prior knoledge. we used grid search and also autoML. we validation by whole one year data (hold year out). we used RMSE

**DS Interviewer Assistant**

Using a hold-out year for validation and RMSE as an evaluation metric is a robust approach.

**Follow-up Question:** Can you discuss why you chose RMSE over other metrics, and what insights it provided you regarding model performance? Additionally, how did you handle potential issues such as outliers in your yield data, and did you apply any data preprocessing techniques before model training?

**User**

outlier is very tricky. we need check with the operation team to make sure it is errors or real cases. we would like to drop extremely cases as accidents always happen in the agriculture.

**DS Interviewer Assistant**

That makes sense; consulting with the operations team is a practical approach to differentiate between genuine outliers and errors.

**Next Question:** Could you describe any other data preprocessing techniques you used apart from outlier removal? For instance, did you perform any scaling, normalization, or handling of missing values before training your models? If so, how did these techniques impact your model's performance?

**User**

exit

**DS Interviewer Assistant**

Thank you for your time and for sharing your experiences. Before we conclude, I'd like to provide you with some feedback regarding your interview performance.

**Evaluation Summary:**

**Score:** 7/10

**Strengths:**

**Technical Depth:** You demonstrated a solid understanding of multiple machine learning algorithms and their application in predictive modeling, particularly in the agricultural domain.

**Problem-Solving Approach:** You highlighted practical strategies for feature selection and handling challenges related to multicollinearity and overfitting.

**Relevant Experience:** Your real-world experiences, especially involving collaboration with the operations team, indicate a practical approach to data science challenges.

**Areas for Improvement:**

**Detail & Clarity:** Some responses could have benefitted from more detail, particularly around data preprocessing techniques and specific outcomes from your model

validation process.

**Advanced Metrics Discussion:** While you mentioned RMSE as your primary metric, discussing additional metrics, such as R-squared or MAE, could have showcased a more comprehensive approach to model evaluation.

Overall, you display strong practical knowledge and problem-solving skills relevant to the role. Thank you for participating in this interview. Would you like to ask me any questions before we finish?

**User**

yes

Enter your message...



Run ⌘ ↵

Playground messages can be viewed by anyone at your organization using the API.