Anticipated Questions

1. How could this project help insurers? This project shows which factors really drive premium costs. By focusing on driver-related variables, insurers can set prices that are both fairer and more precise, reducing risk while keeping customers satisfied.
2. Where did the data come from? It came from a Kaggle dataset that simulates real-world premiums. It includes typical variables insurers use, like age, accidents, and mileage. While it's not proprietary data, it's structured to reflect realistic scenarios.
3. Why do driver details matter more than car details? Because the biggest risk factor in accidents is the human element—experience, age, and history—rather than the vehicle itself. Cars matter, but people matter more.
4. Why start with simple models first? Starting with linear regression gives a clear, interpretable baseline. Once we see how it performs, we can add complexity with models like Random Forest to see if they capture hidden patterns.
5. How well did the models predict? The ensemble models, especially Gradient Boosting, performed best—lowering error rates compared to linear models. That shows the value of non-linear approaches when relationships are more complex.
6. Is it fair to use age in pricing? That's a critical ethical question. Age is predictive, but it can unfairly penalize groups, like young drivers. Insurers must consider fairness alongside accuracy, perhaps by combining age with other, less sensitive variables.
7. What are the weak spots? The dataset is small, and it doesn't include all the real-world variables insurers actually use, like location or vehicle make. That limits generalizability.
8. What new data could improve predictions? Telematics data—things like how often a driver brakes hard, how late they drive, or how often they speed—would make models much stronger and more personalized.
9. How could a company adopt this? Start with a pilot program using historical claims data. Validate that the model works in practice, then expand. It's better to roll out gradually and monitor results than to replace existing systems all at once.
10. Would the results hold on real-world data? Yes, but only if we scale up with larger and more representative datasets. With the right data, these results should generalize well.