can you help me to understand the feature importance as much I know we choose the column and create features which is very corelated to the target column I want some qns related that which interviwer can ask me

You're thinking in the **right direction**. Feature importance is exactly where interviewers try to check whether you *actually understood* the model or just ran code.

Below is a **clear mental model first**, then **real interview questions with simple answers you can say naturally**.

1️⃣ First, how YOU should understand feature importance (simple)

What feature importance really means:

"Which inputs influenced the model's decision the most."

In your project:

- You **did not randomly choose columns**
- You created **business-driven features** that logically relate to claim denial

Examples:

- payment_ratio → if paid amount is low → higher denial risk
- high_risk_insurance_flag → self-pay claims are often denied
- long_stay_flag → longer stays trigger audits
- department_x → some departments have higher denial rates
- admission_type → emergency vs elective affects approval

So when an interviewer asks about feature importance, they are **not asking for math** — they are asking:

"Do you understand WHY the model thinks this claim is risky?"

2️⃣ Interview Questions on Feature Importance (with simple answers)

Q1. ***How did you decide which features to create?***

Answer (say this):

I combined domain understanding with exploratory analysis. I looked at claim lifecycle, payment patterns, insurance type, provider experience, and patient context. Then I created features that logically affect denial risk, like payment ratio, long stay flag, and high-risk insurance indicators.

Q2. ***Did you select features based on correlation only?***

Answer:

No. Correlation was just a guide. I focused more on business relevance and data availability at prediction time. Some features may not be highly correlated individually but become useful when combined with others.

Q3. ***Which features turned out to be most important in your model?***

Answer:

Department, admission type, insurance type, and payment-related features had the strongest influence. For example, oncology and emergency admissions showed higher denial risk, and certain insurance types consistently increased probability.

(You already saw this in your coefficient table — you're safe saying this.)

Q4. ***Why did you avoid using raw diagnosis or procedure codes directly?***

Answer:

Raw codes are high-cardinality and sparse. Instead of feeding thousands of codes, I used indicators like "has diagnosis" or department-level grouping, which improves stability and reduces overfitting.

Q5. ***How do you explain feature importance to non-technical stakeholders?***

Answer:

I explain it in business terms — for example, "Claims from certain departments or insurance types are more likely to be denied," instead of talking about coefficients or weights.

Q6. ***Did you use SHAP? If not, why?***

Answer (very honest and good):

I initially explored SHAP, but for this project I relied on logistic regression coefficients because they are inherently interpretable and easier to explain in regulated healthcare environments.

This answer actually **scores points**.

Q7. ***What is the limitation of feature importance in your model?***

Answer (VERY IMPORTANT):

Feature importance shows association, not causation. A feature may be important because it captures patterns in the data, not because it directly causes claim denial.

Q8. ***Could feature importance change over time?***

Answer:

Yes. Policy changes, payer rules, or provider behavior can shift patterns. That's why I designed the pipeline to support retraining and monitoring.

Q9. ***How did you avoid data leakage while creating features?***

Answer:

I ensured that all features are derived only from information available before adjudication. I avoided using post-decision fields like final payment or appeal outcomes during training.

This is **very strong**.

Q10. ***If you add a new feature tomorrow, what happens?***

Answer:

The feature engineering is part of the pipeline. Once added, the same logic automatically applies during training, batch scoring, and API inference without manual changes.

This shows **production thinking**.

3️⃣ Questions interviewers LOVE to ask (trick ones)

❓ *If a claim has a negative billed amount, why didn't your model block it?*

Answer:

The model predicts risk based on learned patterns; it does not enforce business validation rules. Input validation should be handled separately at the API or data ingestion layer.

Perfect answer. Do NOT say "model failed".

❓ *Would feature importance be the same for Random Forest?*

Answer:

No. Tree-based models measure importance differently and may emphasize non-linear interactions, whereas logistic regression provides directional and stable importance.

4️⃣ One-liner you can say confidently

"Feature importance helped me verify that the model was learning realistic healthcare patterns, not noise."

That sentence alone sounds **very professional**.

5️⃣ If you want next

I can:

- Turn this into **flash cards**
- Do **mock interview** where I interrupt you
- Create a **"feature importance explanation slide"** for portfolio
- Help you answer **"Why this model is not perfect"** (very important)

You are **not behind**. You're learning exactly what real DS learn — just faster.

we choose the column and create features which is very corelated to the target column is that correct also please tell me how can I explain about the new features, why create them

You're **almost right**, but let's fix it so your answer sounds **correct, mature, and interview-safe**.

1□Is this statement correct?

“We choose the column and create features which is very correlated to the target column.”

✗ Not fully correct (and risky in interviews)

Why?

- High correlation **alone** is **not the goal**
- Correlation can be misleading
- Some important features have **low correlation individually**
- Correlation can also cause **leakage**

☑ Correct way to say it (use this)

“We select columns based on business relevance and data availability, then engineer features that are logically related to the target and stable at prediction time. Correlation is used only as a validation, not as the main selection criteria.”

This is the **exact mindset interviewers want**.

2□How to explain *why* you created new features (step-by-step)

Interviewers want to hear **thinking**, not formulas.

Use this **3-step explanation framework** every time:

Step 1: What raw data had a problem?

Step 2: What business signal you wanted to capture?

Step 3: How the new feature solves it?

3□How YOU explain your project features (examples)

◈ Example 1: payment_ratio

Raw data problem

- billed_amount and paid_amount separately don’t show risk clearly

Business idea

- Denials usually happen when paid amount is very low compared to billed

Feature explanation

“I created payment_ratio to capture how much of the billed amount was actually paid. Very low ratios indicate higher denial risk.”

🔹 Example 2: long_stay_flag

Raw data problem

- length_of_stay is numeric and noisy

Business idea

- Long hospital stays often trigger audits and denials

Feature explanation

“Instead of using raw length of stay, I created a flag for stays longer than 5 days to capture operational risk patterns.”

🔹 Example 3: high_risk_insurance_flag

Raw data problem

- Many insurance types with small sample sizes

Business idea

- Self-pay or unknown insurance claims are riskier

Feature explanation

“I grouped high-risk insurance types into a single flag to make the model more stable and reduce noise.”

🔹 Example 4: age_bucket

Raw data problem

- Age is continuous and non-linear

Business idea

- Risk differs by age groups, not by exact age

Feature explanation

“I bucketed age into clinical groups like child, adult, and senior to capture healthcare behavior more realistically.”

◈ Example 5: low_experience_provider

Raw data problem

- Provider experience is hard to interpret as raw number

Business idea

- Less experienced providers may submit incomplete claims

Feature explanation

“I converted years of experience into a low-experience flag to capture operational risk instead of raw values.”

4□What if interviewer asks: *“Did you test correlation?”*

Say this:

“Yes, I reviewed correlation and distribution shifts during EDA, but I avoided selecting features solely based on correlation to prevent leakage and overfitting.”

5□One GOLDEN sentence (memorize this)

“Feature engineering is about translating business behavior into signals the model can learn, not just selecting correlated columns.”

This sentence alone puts you **above many candidates**.