Preguntas teóricas:

1. En la empresa GA, en el área de compras necesitan CLASIFICAR y organizar los correos que llegan a la bandeja de entrada entre 4 tipos de correos (Compras cementos, Compras energía, Compras concretos y correos generales o de otra índole). Esta tarea se le encomienda a usted, gracias a su rol puede solicitar al área interesada los recursos humanos que necesite para llevar a cabo este proyecto, también puede solicitar en tecnología todo lo que necesite, además tiene las bandejas de entrada de correos históricos de los analistas que reciben estas solicitudes con aproximadamente: 5500 correos de compras cementos, 2700 correos de compras de energía, 1100 correos de compras concretos y 12876 correos generales o de otra índole. Explique cómo resolvería este problema, metodología, algoritmos, modelos, arquitectura del proyecto etc.

1. R// Tomamos todos los correos históricos y los clasificamos según las categorías. Cementos, Energía, Concretos y Otros.
2. Preparamos los correos para el modelo, limpiamos el contenido de los correos eliminandos como spam, basura, palabras repetitivas, símbolos, etc para que el sistema pueda leerlo mejor.
3. Hacemos un modelo machine learning, usamos estos correos que ya estab organizados para enseñarle a un programa a identificar de qué tipo es cada correo. Que agrupe por ejemplo las palabras claves, digamos facturas con cemento (Compras cementos).
4. Probamos si aprendió bien, después de entrenar el programa, le damos correos nuevos y revisamos si los clasifica correctamente. Si no lo hace bien, ajustamos el modelo predictivo.
5. Una vez probado y listo, empezamos a ejecutarlo en vivo

En resumen: Básicamente entrenamos un modelo para que clasifique los correos automáticamente y así no tener esa tarea engorrosa manualmente

2. Seis meses después de haber desplegado un modelo de regresión en producción, los usuarios se dan cuenta que las predicciones que este está dando no son tan acertadas, se le encarga a usted que revise que puede estar sucediendo.

R// Es probable que el modelo este sufriendo de Drift, seguramente las condiciones con la que fue entrenado el modelo inicialmente han cambiado. Sucede porque los datos de ahora son diferentes en comparación con los que aprendió el modelo; también puede pasar que la relación entre los datos cambio.

Para validarlo es sencillo, podemos calcular métricas, si estas han empeorado notoriamente, sabremos que hay una dificultad. También podemos hacer una comparación de los datos actuales con los originales, si las distribuciones son muy diferentes podría ser un indicio.

Si hay Drift debemos recolectar nuevos datos, reentrenar el modelo, mantener y mejorar el monitoreo.

3. Su equipo de trabajo está trabajando en un chatbot con generación de texto utilizando el modelo GPT-3.5, según cómo funciona este modelo, ¿cómo haría usted para hacer que las respuestas del chatbot estén siempre relacionadas a conseguir cierta información particular del usuario y no empiece a generar texto aleatorio sobre cualquier tema? Explique su respuesta.

Es importante definir un prompt muy específico, esto nos va a ayudar a limitar ese alcance en las respuestas. Claramente se deben incluir reglas de interacción, darle una lógica externa, si el usuario habla de otro tema guiar nuevamente la conversación al tema central. Y finalmente entrenarlo con datos afinado de manera que aprenda a responder de una forma más ajustada.