Prueba técnica

Teoría:

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.

Entendimiento del Problema

El objetivo es crear sistema eficiente que permita clasificar automáticamente los correos electrónicos entrantes de acuerdo a las cuatro categorías predefinidas con un sistema que debe ser capaz de procesar un gran volumen de correos y asegurar una alta precisión a la hora de clasificación.

Metodología

1. **Tipo de trabajo:** Se estructuraría el proyecto operando bajo una metodología Agíl, dado que, nos permite iterar sobre los inconvenientes que vayan surgiendo y mejorar en pequeños periodos de tiempo el proyecto.
2. **Recolección de datos y almacenamiento:** Se recopilarán todos los correos electrónicos históricos de los analistas y se manejaría como base de datos para el almacenamiento una NoSQL para nuestro caso recomendamos MongoDB para mejorar el procesamiento de la consulta.
   - Otra alternativa puede ser almacenarlas en Azure y generar como trigger power automate.
3. **Análisis exploratorio y limpieza de datos:** Revisar las estructuras de datos de los analistas y realizar una limpieza de los datos para eliminar duplicados, correos dañados, quitar caracteres inválidos y normalizar la información.
4. **Definición de regla de negocio:** Se organizan grupos de trabajo con los expertos del área de compras para poder definir el conjunto de reglas de negocio que se deben aplicar a cada uno de los correos cuando estos lleguen, basándose en el conocimiento y la experticia de cada uno de los expertos del área.

5. **Evaluación y optimización del proyecto:** Se aplicarán métricas como precisión, recall y F1-score: Estas métricas nos permiten ver la exactitud .
   - Ajuste de hyperparametros, nos permitirán mejorar como opera el algoritmo.
6. **Definición e implementación del modelo de aprendizaje automático:** Se utilizaría un modelo supervisado, dado que tenemos las reglas de negocios predefinidas y es fácil de realizar evaluaciones e intervenciones al mismo
7. **Despliegue :** Una vez que el modelo esté entrenado y validado, lo lanzaremos al entorno productivo y revisaremos su comportamiento.
8. **Mantenimiento y mejora continua:** Se realizará una evaluación periódica del sistema para medir su desempeño y detectar posibles problemas.
   Se implementará un mecanismo de retroalimentación con periodicidad mensual para que los usuarios puedan corregir las clasificaciones incorrectas y mejorar así el modelo.

Recursos Necesarios

Humanos:

- Científico de datos
- Ingeniero de software
- Expertos en el área de compras

Tecnológicos:

- Servidor con suficiente capacidad de procesamiento y almacenamiento (Azure)
- Herramientas de desarrollo (Python)
- Librerías de machine learning y NLP (TensorFlow, NLTK)
- Base de datos (MongoDB)

Algoritmos y Modelos Potenciales

- Preprocesamiento: Tokenización, lematización, eliminación de stopwords, vectorización (TF-IDF, word embeddings).
- Clasificación: Naive Bayes, SVM, Random Forest, Redes Neuronales (CNN, RNN, Transformers).
- Clustering: K-means, Hierarchical Clustering. (No Supervisados)

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.

¿Cree que el modelo esté sufriendo Drift?

Si, se tiene alta probabilidad de que este ocurriendo dependiendo del tiempo y si se da concept drift (que las variables subyacentes y variables de salida cambien – ejm: inflación, cambios en la oferta y demanda, etc.) o data drift (Los datos de entrada (features) pueden cambiar con el tiempo – ejm: el comportamiento de compra de los clientes).

¿Cómo puede validarlo?

Para validar si el modelo está sufriendo drift, se pueden usar varias técnicas:

a. Monitoriar el modelo (Métricas de rendimiento): Comenzariamos por analizar si el modelo a empeorado su rendimiento con Error cuadrático medio (MSE) o Error absoluto medio (MAE) para comparar si se ha dado un aumento en el error y aplicar un R2 o coeficiente de determinación, podemos ver con esta métrica si genera una caída, nos indicaría una perdida en la capacidad predictiva.

b. Comprobación de las distribuciones de datos (Data Drift): Validamos los datos de entrada y su cambio de distribución actual, en lo personal utilizo: Histogramas y Boxplots para visualizar las distribuciones actuales de las características de entrada y compararlas con las distribuciones de los datos originales de entrenamiento.

c. Análisis de Concept Drift: revisión de variables subyacentes, lo podemos hacer con un modelo de referencias cruzadas utilizando datos recientes y comprando los resultados con las variables anteriores.

¿De ser así, que haría usted para corregir esto?

Aplicaría los siguientes procesos al modelo:

a. Validación de los datos actuales: Validar que todos los datos utilizados por el modelo sean relevantes y precisos para el momento en que se está realizando el proceso.
b. Selección de características relevantes: Identificar las características que tienen un mayor impacto en la variable objetivo.

c. Validación cruzada: Utilizar técnicas de validación cruzada para evaluar el rendimiento del modelo y detectar posibles problemas de sobreajuste.
d. Reentrenamiento del modelo con datos más recientes
e. Reentrenamiento periódico del modelo de acuerdo con los hallazgos encontrados del análisis de estacionalidad del modelo.
f. Análisis de cambios en las características (Feature Engineering): revisar las características del modelo (tipo de modelo, parámetros, funciones de activación y de perdida utilizadas) y buscar cuales se ajustan más a la realidad.
g. Enfoques de ensemble (combinación de modelos): Utilizar varios modelos en diferentes momentos para combinarlos y generar datos mas robustos.

Posibles técnicas de Ensemble:

- Bagging (Bootstrap Aggregating): Se entrenan múltiples modelos independientes utilizando subconjuntos de los datos de entrenamiento con muestreo aleatorio. Ejemplo: Random Forest.
- Boosting: Se entrenan modelos secuenciales donde cada modelo nuevo intenta corregir los errores de los modelos anteriores. Ejemplo: Gradient Boosting.
- Stacking: Se combinan diferentes modelos entrenados, pero en lugar de hacer votaciones, se entrena un modelo adicional (por ejemplo, regresión logística) para combinar las salidas de los modelos base.

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?

1. Le indicaría un diseño claro de la información que debe recopilar o en que temas debe abordar.
2. Pasarle ejemplos concretos para que el modelo comprenda mejor el contexto.
3. Indicar y explicar al modelo lo que debe evitar y que restricciones debe tener.