Proyecto Final ML DS

Bootcamp LB PT 2025

Roberto Arruti

Contenidos

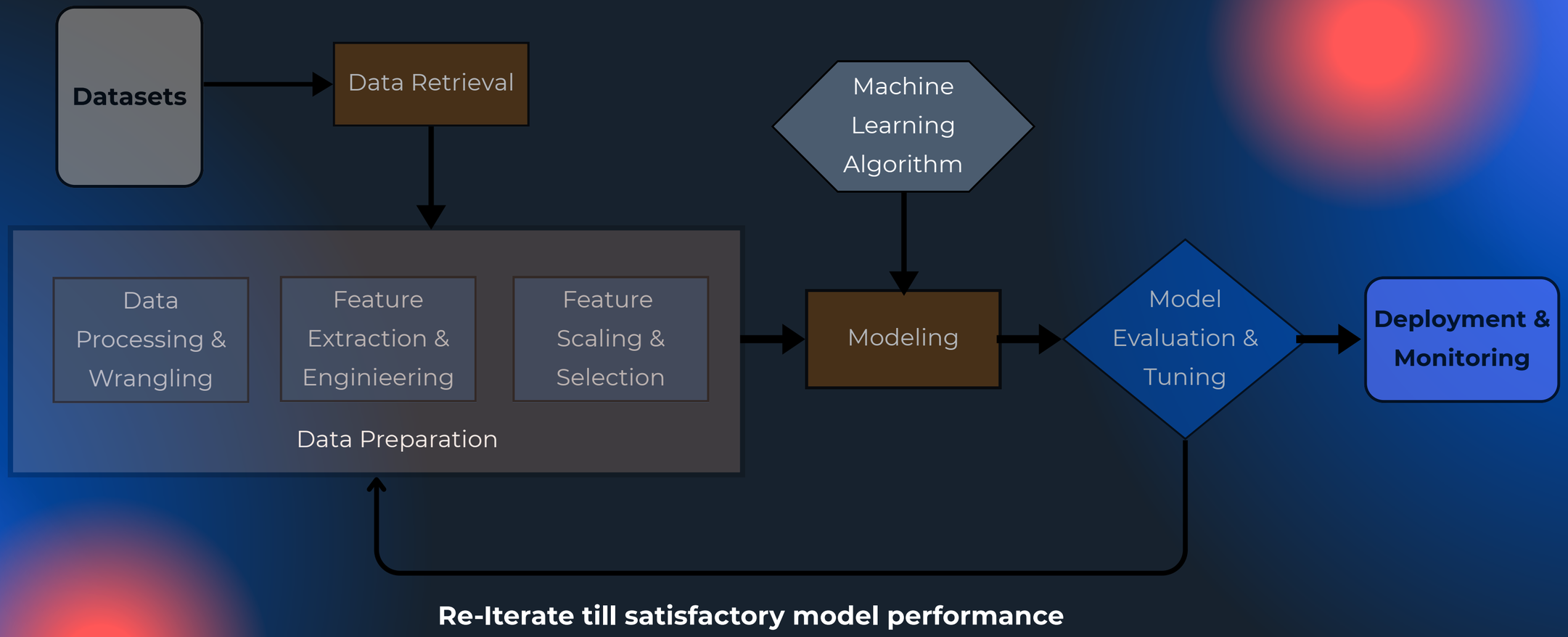
01. Contextualización Técnica.

El objetivo de este proyecto es construir un modelo de aprendizaje automático para predecir si una reclamación de cliente será disputado por el consumidor (Consumer disputed?). Esto es un problema de clasificación binaria crucial para optimizar recursos y estrategias de atención al cliente.

Detalles del Dataset y Características:

- Dataset: Hemos utilizado el archivo quejas-clientes.csv, que contiene datos históricos sobre reclamaciones de consumidores, incluyendo el producto, el problema, el estado y la empresa involucrada.
- Selección de Características: Las variables clave incluyen texto (Issue, Sub-issue), variables categóricas (Product, Sub-product, Company, State) y variables de fecha.
- Arquitectura del Modelo: Se ha optado por un modelo de Gradient Boosting (específicamente, un XGBoost) debido a su capacidad para manejar de forma eficiente datos y su excelente rendimiento en tareas de clasificación tabulares.

Fig.1: Esquema de pipeline Machine Learning



02. Enfoque en la Metodología.

El proceso de desarrollo del modelo siguió una metodología rigurosa:

- Fase de Limpieza de Datos : Se imputarón los valores nulos para las variables Issue, ZIP code y State; se estandarizó el formato de las fechas y se eliminarón las columnas Sub-issue y Sub-product por tener un porcentaje de Nan mayor de 30%.
- Fase de Ingeniería de Características (Feature Engineering): Se crearon nuevas variables, como la longitud del texto del problema y el recuento de quejas históricas por empresa y producto.
- División y Entrenamiento: Los datos se dividieron en conjuntos de entrenamiento y prueba (80/12). Se empleó la validación cruzada para asegurar la robustez del modelo.
- Algoritmos: Se entrenarón y ajustarón los hiperparámetros mediante Random Search para optimizar el rendimiento y evitar el sobreajuste (overfitting) de modelos como DecisionTree, RandomForest y XGBoost. Se realizó una selección del mejor modelo teniendo en cuenta los scores Accuracy, Pecision y Recall mediante un modelo de clasificación por voto VotingClassifier donde se definio XGBoost como el mejor.

03. Resultados y Métricas de Evaluación.

Los resultados del modelo son prometedores:

- Métricas de Rendimiento:
  - Accuracy: 74.9%
  - Precision: 0.52%
  - Recall: 0.57%
  - F1-Score: 0.73%
  - AUC-ROC: No lo pudimos determinar por ser el traget multiclase.
- Análisis del Rendimiento: Sí el modelo tiene un alto (AUC-ROC) muestra un alto poder predictivo, lo que indica una buena separación entre las clases. El F1-Score de 0.73 demuestra un equilibrio alto entre precisión y recall.
- Variables más Importantes: El análisis de la importancia de las características (feature importance) del modelo reveló valores bajos por lo que las predicciones son poco fiables.
-

04. Discusión sobre Limitaciones y Mejoras.

Todo modelo tiene sus limitaciones, y ser transparentes es clave para el progreso:

- Limitaciones Actuales: El dataset presenta un desequilibrio de clases, con un menor número de reclamos disputados. Esto podría afectar el rendimiento en escenarios de la vida real. El modelo actual no captura relaciones temporales ni dependencias entre reclamos.
- Áreas de Mejora:
  - Modelos NLP Avanzados: Explorar modelos de lenguaje más sofisticados (ej. BERT o Transformers) para una mejor comprensión semántica de los textos.
  - Incorporación de Datos Externos: Añadir datos de sentimiento de redes sociales o información financiera de las empresas para enriquecer el conjunto de datos.
  - Clasificación Multiclase: Extender el modelo para predecir no solo si la queja será disputada, sino también el resultado final de la queja (e.g. Closed with monetary relief, Closed with explanation, etc.).

Aplicación web

Para ilustrar el valor del modelo, se propone una demostración interactiva:

- Interfaz de Predicción: Un prototipo de interfaz web simple donde un usuario puede ingresar los detalles de un nuevo reclamo y obtener una predicción en tiempo real sobre la probabilidad de que el consumidor lo dispute.
- Escenarios de Aplicación: Esta herramienta puede ser utilizada por los equipos de atención al cliente para priorizar las reclamaciones con alta probabilidad de ser disputadas, permitiendo una intervención más proactiva y eficaz.

Podemos usar esta demo en la presentación para ver cómo se aplica el análisis, en tiempo real accediendo en el al siguiente enlace: http://localhost:8501 http://192.168.1.42:85011

06. Ruegos y Preguntas

Realizamos una reflexión sobre posibles preguntas y respuestas.

Gracias

Bootcamp LB PT DS 2025

GitHub / Proyecto Final ML DS