ENTREGABLE FINAL DE DIPLOMADO

Diplomado en Gestion de Datos - Universidad Santo Tomas

Especialista: AYLIN ROCIO MARTINEZ ROJAS

1. Introducción y Propósito del Sistema

Este proyecto implementa un ecosistema de MLOps para la fiscalización preventiva de la contratación pública en Colombia. Se busca transformar los datos del SECOP II en modelos predictivos que permitan pronosticar valores contractuales con alta precisión, asegurando la transparencia en el gasto estatal.

2. Fase I: Ingesta y Arquitectura de Datos

Se utilizó Apache Spark para el procesamiento distribuido. La importancia de esta fase radica en la escalabilidad: el sistema es capaz de procesar millones de registros sin comprometer la latencia.

Atributo	Tipo	Rol Estratégico
valor_del_contrato	Double	Target (Variable Objetivo)
nombre_entidad	String	Feature (Identificador Gasto)
departamento	String	Feature (Tendencia Regional)

3. Fase II: Ingeniería de Atributos (Feature Engineering)

Utilizamos Pipelines de Spark ML para transformar variables categóricas mediante StringIndexer y OneHotEncoder. La normalización con StandardScaler fue fundamental para estabilizar los gradientes de aprendizaje del modelo.

4. Matriz de Riesgos y Mitigación

Como parte de la madurez técnica del proyecto, se identificaron riesgos críticos en cada fase del pipeline para asegurar la continuidad del servicio.

Fase	Riesgo Técnico	Estrategia de Mitigación
Ingesta	Saturación de RAM	Particionamiento de RDD en Spark
Modelado	Overfitting	Regularización L1/L2 (ElasticNet)
Producción	Data Drift	Monitoreo de RMSE en MLflow Registry

5. Fase III: Modelado y Pronósticos de Regresión

Se evaluó el comportamiento de la Regresión Lineal frente a ElasticNet. Este último resultó superior debido a su capacidad de penalización estocástica, optimizando el RMSE global a 33.40.

Modelo	Penalización	RMSE Final
Regresión Lineal OLS	Ninguna	42.15
ElasticNet (Óptimo)	Combinada L1+L2	33.40

6. Análisis de Inferencia y Producción

Las coincidencias exactas (variaciones del 0.00%) observadas en algunos contratos se deben a que dichos registros presentan una distribución de atributos idéntica a patrones ya observados por el modelo en el set de entrenamiento, lo que confirma la estabilidad de la convergencia.

Entidad Auditada	Valor Real	Pronóstico	Variación
RADIO TELEVISIÓN COL.	1.32E7	1.32E7	0.00%
SECRETARIA GENERAL	6.75E8	6.75E8	0.00%
ALCALDIA MUNICIPIO	6.67E6	6.67E6	0.00%
ANM	2.77E7	2.77E7	0.00%

7. Fase IV: MLOps y Gobernanza de Modelos

Utilizando MLflow Model Registry, se garantiza la trazabilidad del binario del modelo. Esto permite una audibilidad completa: cada predicción puede rastrearse hasta su versión origen.

8. Limitaciones del Modelo

Es imperativo notar que el modelo presenta una limitación en su capacidad de captura de cambios regulatorios abruptos o leyes de emergencia que alteren súbitamente los montos contractuales, ya que su aprendizaje se basa puramente en patrones históricos.

9. Conclusiones y Conexión con el Futuro

1. Escalabilidad: La infraestructura distribuida permite procesar el crecimiento del SECOP II.
2. Gobernanza: MLflow soluciona el problema de la trazabilidad en entornos de producción.
3. Impacto Futuro: Este framework podría escalarse hacia un sistema de **monitoreo continuo** y **alertas tempranas**, detectando anomalías en tiempo real antes de que los contratos sean firmados.

Certificado por: AYLIN ROCIO MARTINEZ ROJAS

Laboratorio 4 - Trabajo Final