SEGUNDA ENTREGA PROYECTO

XENITAL

Javier Andrés Ramírez Silva - Código 201821781
ja.ramirezs2@uniandes.edu.co

Gabriel Hernández Reyes - Código 201728589
g.hernandezr@uniandes.edu.co

Nicolas Esteban Niño Vega - Código 202324436
n.ninov@uniandes.edu.co

Juan Diego Ospina Aguirre - Código 201814547
jd.ospinaa@uniandes.edu.co

Noviembre 2023

Universidad de los Andes
Ingeniería de Sistemas y Computación
MINE 4101 – Ciencia de Datos Aplicada
Profesores:
Fabian Camilo Peña Lozano
Diego Fernando Ibagón

Punto 1: *[10%] Definición de la problemática y entendimiento del negocio.*

El presente proyecto busca generar competencias mediante la tecnología de la analítica de datos para la compañía XENITAL (Soluciones e Ingeniera) con el fin optimizar sus procesos comerciales ayudando a generar su crecimiento e innovación de oportunidades mejorando sus ventas mediante el ofrecimiento de sus servicios, el cual permitirá realizar un análisis profundizado de los datos empleando diferentes herramientas para su análisis descriptivo, predictivo y prescriptivo para la obtención de información privilegiada para la toma de decisiones futuras.

XENITAL es una compañía colombiana con sedes en Bogotá, Cali y Barranquilla, la cual presta servicios para ingeniería, realizando adquisición de información espacial, como: Información y datos en el sector de la ingeniería de vías y transportes, además, busca implementar tecnologías de captura de datos, procesamiento y análisis, para aumentar la competitividad en los clientes y usuarios de una manera rápida, práctica, económica y eficiente.

Problemática:

Actualmente existe una alta demanda en el mercado acerca de la implementación de nuevas tecnologías en la rama de la ingeniería. El uso de drones en las operaciones comerciales se ha incrementado en diferentes industrias en los últimos años debido a su capacidad de mejorar la eficiencia y el análisis de datos para la reconstrucción tridimensional de diferentes proyectos, terrenos o topografías.

Dentro de este contexto, existe una gran oportunidad para mejorar la adquisición de información y optimizar los procesos, mediante el pronóstico del tiempo de vuelo de los drones y posteriormente la identificación de la cantidad de pilotos requeridos para la operación, lo cual le permite a Xenital hacer una planeación acertada del recurso humano necesario para cumplir y fortalecer la eficiencia operativa para la alta demanda, así mismo mejorar con la toma de mejores decisiones y proyectar el presupuesto requerido para la operación. Dentro de los principales beneficios, se encuentran:

- Agilidad en los procesos de contratación o desvinculación.
- Aumento de la eficiencia operativa.
- Planeación y ejecución del presupuesto.
- Planeación acertada del recurso humano para la demanda.
- Mejorar calidad disminuyendo costos en un menor tiempo.
- Alcanzar una posición competitiva.

En conversación con el negocio, se requiere pronosticar el tiempo de vuelo de los drones en los próximos meses y predecir el nivel de consumo de la batería con el fin de planificar distintos procesos dentro de la compañía con anticipación, como adelantar procesos de contratación o desvinculación de pilotos, conocer el presupuesto requerido para el pago de pilotos y aumentar la eficiencia operativa al tener a la cantidad de pilotos necesarios para la demanda mensual.

II. Entendimiento del Negocio

Servicios

Fotogrametría con dron (RPAS): Consiste en garantizar una atención profesional y personalizada acorde con las necesidades del cliente, y con la orientación requerida para la correcta toma de decesiones de manera oportuna y acertada, mediante el seguimiento físico y financiero de una o varias actividades desarrolladas en un proyecto, empleando equipos tecnológicos idóneos para la captura y recopilación de información que permita desarrollar un producto satisfactorio para la compañía y para el cliente.

- Seguimiento e Inspección de Obras.
- Localización y fotos 360°.
- Ortomosaico General.
- Inspección Indoor O Interna.
- Planeación

Financiero

Dentro del estudio de oferta para levantamientos topográficos dentro del mercado actual, la obtención de precios fijos, variables, precios unitarios por hectárea tanto para la venta como el costo de producción, punto de equilibrio el cual permite comprobar la viabilidad del negocio, el nivel de beneficios y además es uno de los elementos centrales en la constitución de cualquier tipo de empresa, pues es clave a la hora de determinar la solvencia de la empresa y su nivel de rentabilidad.

III. Descripción del sector:

El sector de la ingeniería civil amplía sus posibilidades y potencial con el uso de una herramienta de autonomía excepcional como los drones. Mediante procesadores, software y GPS aportan a las empresas información espacial importante a través de sensores incorporados en las aeronaves. Las posibilidades son infinitas para ingeniería civil, ya sea en inspecciones de construcción, levantamientos topográficos, mantenimiento de infraestructuras energéticas, impacto medioambiental, imágenes térmicas exploración geológica, gestión de riesgos, etc. Existen alrededor de 165 empresas vinculadas al sector, ubicadas en todo el territorio nacional.

Punto 2: *[10%] Ideación.*

El potencial usuario del producto sería el COO - Chief Operating Officer, encargado del área operativa, lo que incluye procesos de contratación o desvinculación, eficiencia operativa, planeación y ejecución del presupuesto, planeación acertada del recurso humano para la demanda, entre otras. A través de la recolección de datos topográficos y gastos en los recorridos, tanto de operadores como de operación de los drones, se requiere que el producto permita identificar el grado de confiabilidad de inversión y rentabilidad de los proyectos.

Los requerimientos del producto de datos como los componentes desde un punto de vista tecnológico además de la recolección, procesamiento y limpieza de los datos,

son el almacenamiento, escalabilidad y visualización de los datos, como el uso de API y servicios en la nube.

Punto 3: *[10%] Responsable.*

Teniendo en cuenta que XENITAL ofrece el servicio de seguimiento e inspección de obras, se documenta el estado del arte referente al caso con una de sus principales herramientas como “FLY.AI”, la cual es una plataforma escalable y abierta de procesamiento automático mediante inteligencia artificial que permite la incorporación de nuevas infraestructuras, drones de última generación, sensores, imágenes de satélite y modelos de IA propios o de terceros, cubriendo todas las etapas del ciclo de vida de una inspección:

- Planificación de la misión (calendario, zona a revisar, pilotos y drones).
- Registro de documentación exigida por la normativa vigente en cada país.
- Evaluación del riesgo y aprobación.
- Ejecución del vuelo y registro en plataforma de todos los datos o imágenes capturados.
- Generación automática de informe de incidencias (empleando los modelos de IA integrados).
- Exportación del informe en diferentes formatos para su entrega a terceros.
- Visualización del mapa de la infraestructura.

La base de una buena inspección es una excelente toma de datos. Una de las características más innovadoras que presenta FLY.AI es: una vez subidos los archivos a la plataforma, estos se pueden analizar de forma desatendida y en cuestión de minutos se obtienen los resultados del algoritmo de Inteligencia Artificial.

Así mismo el grado de adecuación de los datos recabados a los intereses de XENITAL respecto al enfoque de la explotación de los mismos, su carácter completo o incompleto, su veracidad, fiabilidad y su homogeneidad. Es decir, la calidad de los datos se trasvasa a las informaciones que se obtienen a partir de ellos, clave para una toma de decisiones acertada con el fin de **definir una estrategia** de gestión de datos y gobernanza de la información.

Factores básicos con el fin de establecer el nivel de calidad de los datos:

INTEGRIDAD: Los datos sean completos, se eviten duplicidades y se dispongan las medidas necesarias para evitar cruces e interferencias.

OPERATIVIDAD: Los datos sean lo suficientemente homogéneos, sólidos y consistentes para permitir una explotación adecuada de los mismos.

VERACIDAD: Disposición de un comprobado valor referencial.

De esta manera se estable el uso de técnicas de análisis de datos para mejorar la profundidad y la cuantificación de los problemas y así llegar a una identificación más oportuna de los riesgos presentes y futuros basado en muestras a monitoreo continuo analítico basado en revisiones de la totalidad de los datos.

Se considera realizarse un análisis por cada una de las dimensiones de calidad de datos, logrando de esta manera resolver cada una de las dudas existentes en el proceso y mitigando así los riesgos teniendo un punto de partida, una métrica que permita identificar el estado actual de los datos. Para ello, es básico realizar una auditoría inicial o perfilamiento de los datos, con el objetivo de averiguar en qué estado se encuentran y a su vez determinar parámetros de control que ayuden a medir el avance en los procesos de calidad.

Punto 4: *[15%] Enfoque analítico.*

Escenarios:

I. **Pronosticar el tiempo de vuelo máximo (autonomía del dron) asumiendo un consumo alto de recursos.**

Hipótesis: *¿Qué tipo de tareas puedo permitirme hacer con el tiempo de vuelo máximo ofrecido con un consumo alto de recursos?*

Alineación estratégica y de negocio: Fortalecer la productividad y competitividad en el sector mediante mejores prácticas contribuyendo a un portafolio de servicios más competitivo en el mercado mediante la eficiencia operativa.

- Tipo de aprendizaje: Supervisado.
- Tarea de aprendizaje: Regresión.
- Técnica / Algoritmo: Regresión Lineal – Regresión Logística.

Al ser un problema de regresión en el que se utilizarán algoritmos de machine learning, las métricas que se emplearán para la medición de los modelos serán: MSE, RMSE, MAE, R2, entre otros.

II. **Predecir el nivel de consumo de la batería con el fin de identificar si el tipo de vuelo es de consumo bajo, medio o alto consumo para prever que se agoten.**

Hipótesis: *¿Se tiene identificado el nivel de consumo de batería del vuelo que queremos realizar para garantizar una mejor eficiencia operativa?*

Alineación estratégica y de negocio: Evitar horas de vuelo perdidas sin lograr el objetivo del vuelo por falta de predicción en el consumo promedio del vuelo. Esto con el fin de ser una compañía competitiva en el mercado mediante la implementación de soluciones tecnológicas basadas en datos, contribuyendo a la investigación, el desarrollo y la integración de diversas tecnologías para el control de activos.

- Tipo de aprendizaje: Supervisado.
- Tarea de aprendizaje: Clasificación.
- Técnica / Algoritmo: K-Nearest Neighbors – Árboles de decision – Random Forrest – SVM.

Al ser un problema de clasificación en el que se utilizarán algoritmos de machine learning, las métricas que se emplearán para la medición de los modelos serán la exactitud, la precisión, recuperación (recall), puntuación F1 (F1 score), área bajo la curva ROC, entre otros.

Punto 5: *[10%] Recolección de datos.*

Por facilidad y por la necesidad de la ejecución de scripts en Python para solucionar este requerimiento, este punto se encuentra en el notebook que se encuentra en el repositorio.

Punto 6: *[35%] Entendimiento de los datos.*

Por facilidad y por la necesidad de la ejecución de scripts en Python para solucionar este requerimiento, este punto se encuentra en el notebook que se encuentra en el repositorio.

Punto 7: *[10%] Primeras conclusiones, insights y acciones próximas a ser ejecutadas.*

Este punto se encuentra en el notebook que se encuentra en el repositorio.

II SEGUNDA ENTREGA

De acuerdo al alcance definido en el presente proyecto donde se busca generar competencias mediante la tecnología de la analítica de datos para la compañía XENITAL (Soluciones e Ingeniera) con el fin optimizar sus procesos comerciales ayudando a generar su crecimiento e innovación.

Dentro de este contexto, mediante el pronóstico del tiempo de vuelo de los drones y posteriormente predecir el nivel de consumo de batería del dron, lo cual le permite a Xenital hacer una planeación acertada del recurso humano necesario para cumplir y fortalecer la eficiencia operativa, así mismo mejorar con la toma de mejores decisiones y proyectar el presupuesto requerido para la operación.

Punto 8: *[35%] Preparación de datos.*

Para la preparación de los datos se realizaron las siguientes actividades previas:

- Exploración y Entendimiento del conjunto de los datos

En primera instancia se quiere entender los datos que proporciona la empresa:

Se compartieron datos de 3000 vuelos de Xenital, con 22 columnas, de las cuales 11 son de tipo numérico y el restante (11) son de tipo categórica.

- Teniendo en cuenta que la técnica inicial plateada fue regresión y clasificación donde se definieron los siguientes casos:

Caso 1: Pronosticar el tiempo de vuelo máximo (autonomía del dron) asumiendo un consumo alto de recursos. (Regresión)

Variable Objetivo: *Air Seconds*

Caso 2: Predecir el nivel de consumo de la batería con el fin de identificar si el tipo de vuelo es de consumo bajo, medio o alto consumo para prever que no se agoten. (Clasificación)

Variable objetivo: *Differece Bat %, donde:*

- Si Difference Bat % >= 66.67% => Consumo alto.

- Si Difference Bat % > 33.33% ^ Difference Bat % < 66.67% => Consumo medio.
- Si Difference Bat % <= 33.33% => Consumo bajo.

- Perfilamiento de datos (Selección de variables)

Previo al perfilamiento de datos, se realizó la única modificación al dataset donde se eliminaron las variables que no fueron relevantes para los análisis, así.

Variable	Descripción
Pilot-in-Command	La experticia piloto es una variable influyente en las métricas de vuelo.
Air Seconds	Variable objetivo - tiempo de vuelo influye directamente en el consumo de batería
Above Sea Level (Meters)	Presión atmosférica (debido a la altura) puede afectar significativamente el desempeño del vuelo.
Drone Type	Variable principal - aspectos técnicos de la aeronave, donde dependen variables (como batería y otros)
Takeoff Bat %	Consumo de batería
Takeoff mAh	
Takeoff Volts	
Max Altitude (Meters)	Variable relevante - la altitud alcanzada influye en el consumo adicional de batería
Total Mileage (Kilometers)	La distancia total es directamente proporcional al consumo de la batería

Se seleccionaron estas variables debido a que es posible conocer su valor a priori (antes del vuelo). Para el caso de Max Altitude (Altura máxima) y Total Mileage (Kilometers) el usuario debe realizar una estimación de cuanta altitud espera alcanzar en el vuelo, así como la distancia total que espera recorrer.

- Análisis y corrección de valores nulos

Se identificó que hay 3 campos con valores nulos, de los cuales 1 tiene un porcentaje mayor al 30%, para este caso es impensable realizar algún proceso de completado dado que es un porcentaje demasiado alto, por lo que se opta por eliminar esta columna.

Ahora bien, para los 2 campos restantes con valores nulos, se realizará una imputación de datos, Dado que las distribuciones no son completamente normales, se hará la imputación de los datos por la mediana.

- Codificación de variables categóricas

Se realizó la codificación de las variables consideradas categóricas (**Pilot-in-Command y Drone Type**) con el fin de buscar una estandarización de datos, evitando que tengas características individuales o no contantes de acuerdo a su desviación estándar.

- Normalización y estandarización de los datos

Bajo el empleo de la librería Scikit-learn se obtuvieron los valores normalizados de los atributos numéricos representando novedades en las escalas, para ellos se restó la media de los valores y se divido por la desviación estándar de los mismos.

A si mismo se identificaron valores atípicos u/o outliers obtenidos por la normalización y seguidamente haciendo empleo del método (fit) se calculó la media y la desviación estándar de los valores.

Finalmente, con el método (transform), se obtuvo los valores ya escalados. Dando un ranquin de valores ya escalados imprimiéndolos.

El presente punto se encuentra con mayor detalle en el notebook, ubicado en el repositorio.

Punto 9: *[10%] Estrategia de validación y selección de modelo.*

Para los cos casos se entrenarán varios modelos, unos de regresión (Caso 1) y otros de clasificación (Caso 2), además se utilizará GridSearch de Scikit-learn para encontrar los mejores hiperparámetros del modelo, esto con el conjunto de entrenamiento.

Caso 1

- Regresión Lineal
- Regresión Lineal con regularización L1 (Ridge)
- Regresión Lineal con regularización L2 (Lasso)
- Regresión con Random Forest
- Regresión con Gradient Boosting
- Regresión con Máquinas de Soporte Vectorial (SVR)

Caso 2

- Clasificador Random Forrest.
- Clasificador Gradient Boosting.
- Clasificador Máquinas de Soporte Vectorial (SVM).
- Regresión Logística.

Posteriormente, para los dos casos se evaluarán los modelos con los hiperparámetros encontrados. Esto se realizará con el conjunto de prueba. A partir de las métricas encontradas se escogerá el mejor modelo. Este mejor modelo se validará con los datos del conjunto de validación, encontrando las métricas del mejor modelo encontrado.

Punto 10*. [20%] Construcción del modelo.*

El entrenamiento y desarrollo de los modelos utilizados para predecir los dos casos propuestos se encuentran el respectivo notebook, ubicado en el repositorio.

Punto 11 *[20%] Evaluación del modelo.*

Caso 1:

La comparación de los resultados obtenidos mediante diversos algoritmos revela que el modelo de Regresión Random Forest se ha destacado como una herramienta sólida en la estimación del tiempo de vuelo, expresado en segundos.

En el contexto del conjunto de entrenamiento, este modelo exhibe una impresionante capacidad de ajuste a los datos, logrando capturar el 97.2% de la variabilidad presente. Como es una tendencia habitual en numerosos modelos, se experimenta cierta disminución en la precisión al aplicarlos a los conjuntos de prueba y validación, con un 78.6% y un 82.3% de variabilidad explicada, respectivamente. No obstante, es pertinente subrayar que el modelo ostenta un nivel de ajuste adecuado. Con respecto a las métricas de error, se presentan los siguientes resultados:

- ✓ Coeficiente de Determinación (R2)=0.79
- ✓ Error Cuadrático Medio (MSE)=34719.94
- ✓ Error Absoluto Medio (MAE)=126.11
- ✓ Raíz del Error Cuadrático Medio (RMSE)=186.33

Caso 2:

El modelo de clasificación confeccionado mediante la implementación del Random Forest Classifier ha demostrado un desempeño meritorio en términos de métricas de precisión, Recall y puntaje F1, específicamente en el contexto de las categorías correspondientes al bajo y alto consumo de energía en el dron. No obstante, se ha enfrentado a desafíos notables en la correcta categorización de los casos de consumo intermedio, lo que señala la particular eficacia del modelo en la predicción de los extremos del espectro de consumo de batería, pero sus limitaciones en cuanto a la precisión en situaciones de consumo medio.

En el conjunto de prueba, el modelo ha exhibido una precisión global aceptable del 80%. Es relevante destacar que, aunque la precisión para la categoría de consumo medio ha sido relativamente baja (57%), el modelo ha conseguido un equilibrio favorable entre precisión y Recall para las categorías de bajo y alto consumo. En un escenario de validación, el modelo también ha alcanzado un nivel de precisión del 80%, con una precisión excepcionalmente alta del 93% para la categoría de bajo consumo y una precisión respetable del 78% para la categoría de alto consumo. No obstante, es esencial señalar que la precisión para la categoría de consumo medio se ha visto notoriamente reducida (60%).

El desarrollo del presente punto se encuentra con mayor detalle en el notebook, ubicado en el repositorio.

Punto 12 *[15%] Conclusiones.*

¿Cuáles son las mayores dificultades que se han tenido en el proyecto?

En la etapa de calidad de datos la consistencia y completitud de datos ha tenido un mayor impacto en el modelo lo cual hizo generar distintas estrategias en su respectiva etapa, la ausencia de datos puede ser un factor importante en la predicción de nuestras variables objetivos. Aunque en el caso 1 se ha podido predecir de forma muy buena el tiempo de vuelo del dron, para el caso 2 el resultado preliminar se puede mejorar con la inserción de dicha información.

El no tener un pleno conocimiento sobre el tema desarrollado dificulta n análisis más preciso con base a la aplicación del modelo en campo, ya que es posible que se estén obviando características importantes en el desarrollo del modelo.

a. *¿Qué estrategias se plantean para mitigarlas?*

I. Solicitar una mayor cantidad de datos con el fin que el modelo tiene la oportunidad de aprender patrones más complejos y sutiles en los datos, mejorando la estabilidad y aumentando la capacidad de descubrir características relevantes.
II. Probar otros algoritmos adecuados y creación de modelos, entrenarlos (sin sobre entrenarlos) y validarlos, y por último probarlos con el fin de poder identificar otros posibles modelos.
III. Realizar mejas conjuntas con el personal de la empresa Xenital con el fin de poder conocer más el contexto de las actividades desarrolladas en el contexto de negocio y poder tener un panorama mucho más completo.

b. *¿Qué condiciones considera que deberían tener los datos para obtener mejores resultados? Más datos, diferentes características, menor sesgo, etc.*

I. Introducir nuevas variables en un contexto mucho más cercano a la realidad (clima, velocidad del viento, etc.) que puedan hacer que el modelo se pueda desarrollar de una forma menos teórica y más práctica.
II. Que tengan calidad, los datos deben ser precisos y confiables es decir no deben contener errores, valores atípicos o información incorrecta.
III. Que sean completos, los datos deben ser completos, es decir no deben existir valores faltantes o nulos en las variables críticas.

c. *¿El mejor modelo obtenido hasta el momento es suficiente para dar solución al problema u oportunidad de negocio abordado?*

En el primer escenario, los análisis iniciales arrojaron resultados prometedores, lo que sugiere que, en la próxima etapa del proyecto (entrega 3), con ajustes adicionales, como la implementación de métodos para evitar el sobreajuste y la posible incorporación de datos adicionales para el entrenamiento, se espera que el modelo mejore su precisión y confiabilidad en situaciones del mundo real. Esto lo convertiría en una herramienta valiosa para la planificación y gestión de vuelos de drones.

En el segundo caso, el modelo ha demostrado su utilidad al clasificar eficazmente los extremos del consumo de batería en el dron. No obstante, requiere mejoras para realizar predicciones precisas en situaciones de consumo intermedio. Sería beneficioso explorar características adicionales o considerar estrategias avanzadas de ingeniería de características con el fin de potenciar la capacidad del modelo para distinguir entre las categorías intermedias de consumo.