***Objetivos del Proyecto Entrega Plus***:

El Objetivo del proyecto es poder crear un modelo que pueda identificar para cada pedido, si el mismo se va a entregar o no en la primera visita que realice nuestro proveedor logístico.

**Aclaración:** el proyecto tiene 3 etapas en total, pero nuestro alcance implica participación muy activa en las primeras 2:

* **ETAPA 1:** Estudio de los datos
* **ETAPA 2:** Desarrollo del Modelo de Machine Learning
* **ETAPA 3:** Integración con proveedores Logísticos

***Contexto Proceso de Distribución:***

Comienza apenas un distribuidor retira de los Centros de Preparados, los pedidos que le han sido asignados. En este proceso, cada pedido puede tener los siguientes estados:

* *En Curso:* el pedido se encuentra en poder del distribuidor
* *Entregado:* el pedido ha sido entregado al cliente
* *Devuelto:* el pedido ha sido devuelto al centro que preparó el pedido
* *Siniestrado:* el pedido ha sufrido un robo o hurto
* *Destruido:* los pedidos en los que solamente se solicitan "Sims", el proveedor si no los puede entregar al cliente, se encarga de la destrucción de este tipo de material

***"Epsilon S.A."*** nos informa varias problemáticas a abordar, y nos proponen los siguientes proyectos (a desarrollar por etapas), porque se quieren apalancar de la Ciencia de Datos para mejorar sus procesos:

* Predicción de Ingreso de Stock
* Predicción de la demanda de Pedidos (próxima semana, próximo mes, próximo año)
* Mejora de la Efectividad de la Primer Visita del Proveedor Logístico

Según el relevamiento realiza en conjunto, se ha decidido avanzar con el proyecto de ***"Mejora de la Efectividad de la Primer Visita del Proveedor Logístico"***, por las siguientes razones:

* Mejorar la imagen ante nuestros ***CLIENTES***
* Evitar reclamos teléfónicos/por mail al área comercial, por parte de los clientes, por demoras en el arribo de sus pedidos
* Reducir costos de distribución, ya que, mejorando la tasa de entrega en la primera visita, vamos a:
  + Evitar segundas o terceras visitas, lo cual trae un costo extra aparejado
  + Evitar Devoluciones innecesarias de pedidos

***Hipótesis:***

***Primer Hipótesis***: La zona geográfica donde se envían los pedidos influye en nuestros indicadores de:

* Puntualidad de Entrega de Pedidos
* Efectividad de Entrega de Pedidos

Con lo cual entendemos que sabiendo donde debemos entregar cada uno de ellos, podemos establecer una probabilidad de ***NO ENTREGA***, que nos va a permitir tomar acciones para que cada pedido con baja probabilidad de entrega, sea finalmente entregado en su primer visita

***Segunda Hipótesis:*** Los proveedores logísticos pueden ser más fuertes que su competidor en una determinada zona, y el otro, en otra

***Tercer Hipótesis:*** Mejorar nuestros indicadores implica no solo una excelencia para la experiencia global del cliente, sino que también un ahorro enorme a nivel presupuestario

***Data Acquisition***

***Metadata:***

* ***Descripción de los datos:*** El set de datos es acerca de pedidos de la empresa ***EPSILON S.A.***, los cuales hacen referencia a los detalles para la entrega de los mismos a nuestros clientes.
* **Detalles generales del dataset**
  + Cantidad de Registros iniciales: 190152
  + Cantidad de Registros finales: 189834
  + Volumen de Crecimiento Estimado Diario: 3500/4000 (incluye otros países y todos sus canales de venta)
  + Datos del año 2024
* **Diccionario de datos**

|  |  |  |
| --- | --- | --- |
| **Columna** | **Tipo de Dato** | **Tipo de Variable** |
| Entrega | int64 | Clave Primeria (index) |
| Fecha Pedido | datetime64 | Temporal |
| Distribuidor | category | Categórica Nominal |
| Distribuidor Sigla | category | Categórica Nominal |
| Tipo de Distribución Comercial | category | Categórica Nominal |
| Codigo Postal | int64 | Categórica Nominal |
| Id Provincia | int64 | Categórica Nominal |
| Provincia | category | Categórica Nominal |
| Localidad | category | Categórica Nominal |
| Departamento | category | Categórica Nominal |
| Aglomerado | category | Categórica Nominal |
| Id Zona Logistica | int64 | Categórica Nominal |
| Zona Logistica | category | Categórica Nominal |
| Zona E-Commerce | category | Categórica Nominal |
| Periodo | category | Categórica Nominal |
| Año | int64 | Categórica Nominal |
| Mes Sigla | category | Categórica Nominal |
| Mes Nombre | category | Categórica Nominal |
| Mes Numero | int64 | Categórica Nominal |
| Trimestre Sigla | category | Categórica Nominal |
| Trimestre Nombre | category | Categórica Nominal |
| Trimestre Numero | int64 | Categórica Nominal |
| Semestre | int64 | Categórica Nominal |
| Año Semana | category | Categórica Nominal |
| Dia Semana Sigla | category | Categórica Nominal |
| Dia Semana Numero | int64 | Categórica Nominal |
| Dia Semana Nombre | category | Categórica Nominal |
| Hora Pedido | int64 | Categórica Nominal |
| Visita 1 Fecha | datetime64[ns] | Temporal |
| Visita 1 Hora | object | Temporal |
| ID Visita 1 Motivo | category | Categórica Nominal |
| Visita 1 Motivo | category | Categórica Nominal |
| Dias Primer Visita | int64 | Categórica Ordinal |
| Flag Visita 1 Puntual | category | Categórica Nominal |
| Estado Visita | category | Categórica Nominal |
| Estado Distribución | category | Categórica Nominal |
| Calidad de la Direccion | category | Categórica Nominal |
| Cantidad de Pedidos | int64 | Categórica Nominal |
| Visita 1 Puntual | int64 | Categórica Nominal |
| Cant. Primera Visita | int64 | Categórica Nominal |
| Cant. Vis. y Ent. 1ra Visita | int64 | Categórica Nominal |
| Cantidad Celulares | int64 | Numérica Discreta |
| Cantidad Accesorios | int64 | Numérica Discreta |
| Peso Bruto | float64 | Numérica Contínua |
| Valor Total en Pesos | float64 | Numérica Contínua |
| Valor Total en Dolares | float64 | Numérica Contínua |
| Provincia Sigla | object | Categórica Nominal |
| Hora Visita | int64 | Categórica Nominal |

***Gráficas Importantes:***

Gráficas para analizar las ***Hipótesis 1*** e ***Hipótesis 2***:

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Tal como se observa en ambas gráficas, hay zonas del país en donde se observa que la ***“Efectividad de Entrega”*** de las mismas es superior a otras. Por ejemplo, la zona MEDI.

También se observa lo contrario, hay zonas en las que se observa que la ***“Efectividad de Entrega”*** baja muchísimo independientemente del proveedor.

También observamos que ***“Correo Andreani”*** tiene un desempeño superior a ***“Correo Argentino”*** en cuanto a la ***“Efectividad de Entrega”*** a nivel nacional. Pero también se ve claramente lo planteado en la hipótesis. Hay proveedores más fuertes que otros en la misma zona. Ejemplo: ***“Correo Argentino”*** en ***GBA***.

La tercera hipótesis, se va a poder comprobar luego de la implementación del proyecto.

***Data Wrangling:***

* Se han concatenado Dataframes
* Se ha hecho merge entre el dataset principal (Pedidos) y otro obtenido de una API pública con datos puntuales de cada provincia
* Se trataron registros en NaN
* Se ha modificado el tipo de datos de las columnas
* Se han ajustado los nombres de las columnas
* Se han creado nuevos campos
* Se han borrado los campos que no resultaban necesarios para el análisis
* Se ha creado el índice del dataset principal

***Análisis Exploratorio de Datos:***

Se ha realizado un análisis preliminar, en el que se informó que:

* No existen valores duplicados ya que el origen de datos nos asegura que, por su propia restricción de clave, que esa situación no va a ocurrir. Cabe aclarar que, para este Dataset, en el sistema de origen, la clave/Id es el número de Entrega
* El Numeral (#), en el sistema de origen de los datos, representa aquellos valores nulos. Por lo visto en este análisis preliminar, los únicos campos con datos nulos son "Visita 1 Fecha" y "Hora Visita". Es perfecto que eso así sea ya que son pedidos que aún no han sido visitados
* Los valores de 0 (cero) en los campos del Dataset, son coherentes al 100% con el valor que representan
* El siguiente Gráfico demuestra la gran calidad del dataset

Gráfico

Descripción generada automáticamente con confianza media

Se ha estudiado la cantidad de pedidos por zonas:

Gráfico, Gráfico de barras

Descripción generada automáticamente

Como se observa en la gráfica, hay zonas donde la demanda de pedidos es superior a otras, en casi el doble de unidades:

* + Buenos Aires y La Pampa (BALA)
  + Gran Buenos Aires (GBA)
  + Misiones, Entre Ríos y Corrientes (MEDI)
  + El resto de las zonas, con valores similares
  + El estudiar cada zona por separado nos puede dar información de valor para saber cuáles de ellas se deben abordar con mayor prioridad

Se ha estudiado la distribución de pedidos por distribuidor:

Gráfico, Gráfico circular

Descripción generada automáticamente

* Correo Andreani tiene el 90% de la distribución contra el 10% de Correo Agentino
* Pero hemos visto en una gráfica anterior, el buen desempeño de Correo Argentino en GBA (Gran Buenos Aires). Con lo cual este análisis puede ayudar a que las áreas comerciales decidan si quieren aumentar la cantidad de pedidos diarios a este correo en esa zona
* Del mismo modo, quizás se deba evaluar por ese mismo análisis, sacar distribución a ese correo en zonas como AMBA, Litoral Sur y Noroeste Argentino

***Análisis Estadistico de Datos:***

* Se ha estudiado la cantidad total de pedidos por mes:

Gráfico, Gráfico de líneas

Descripción generada automáticamente

* A nivel estadístico, para este análisis específico de estadísticas, se ven datos razonables de:
  + Mínimo (11067)
  + Máximo (24854)
  + Media (17257)
  + Media Geométrica (16775)
  + Media Armónica (16321)
  + Media Recortada (17101)
* La moda que aporta el método pierde valor ya que no hay repeticiones en el conjunto de datos
* Se observa una muy alta varianza (19248473) pero contrariamente, un bajo nivel del coeficiente de variación.
* El desvío estándar razonable con la unidad de cantidad de pedidos mensuales respecto a su media (4387)
* También la gráfica demuestra tener una asimetría positiva según SKEWNESS (0.59)
* Hay exceso de Kurtosis ya que la misma es negativa (-0.71)

***Análisis Univariado:***

Las mayoría de las distribuciones tienen variables categóricas, más allá de que sean números enteros ("Trimestre Número", "Mes Numero", etc.)

Las que analizamos porque son representativas son "Dias Primera Visita", Peso Bruto", "Valor Total en Pesos" y "Valor Total en Dolares"

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Gráfico, Gráfico de barras

Descripción generada automáticamente

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

En las gráficas anteriores se estudió el “Peso Bruto”, “Valor en Pesos” y “Valor en Dolares” por provincia. Se ven claramente los outliers en todas las gráficas. Los mismos fueron tratados mediante asignación del valor medio.

Imagen que contiene Interfaz de usuario gráfica

Descripción generada automáticamente

Escala de tiempo

Descripción generada automáticamente

Gráfico

Descripción generada automáticamente

En las gráficas anteriores se volvió a estudiar el “Peso Bruto”, “Valor en Pesos” y “Valor en Dolares” pero ahora por Distribuidor. Se ven claramente los outliers en todas las gráficas como era de esperar, pero ahora tenemos el dato de que Correo Andreani distribuye mayor cantidad de pedidos en esa situación.

***Análisis Bivariado:***

Para este análisis hemos analizado la “Valor en Pesos” y “Valor en Dolares” pero con la variante de que las esferas tienen el tamaño del “Peso Bruto”

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Se ve claramente una alta correlación de las variables, y también un tamaño de esferas variados para ambos Distribuidores.

Al ser una correlación tan alta, también concluimos que no es necesario tener ambas variables en el modelo porque ambas explican lo mismo.

También se han analizado las matrices de correlación. Por ejemplo, Pearson:

Gráfico

Descripción generada automáticamente

* Se observa alta correlación entre las distintas clasificaciones de fecha ("Mes Numero", "Trimestre Número" y "Semestre").
* Se observa también una muy fuerte correlación entre el valor de cada "Entrega" en Pesos Argentinos y Dólares. Esto si se puede ver como algo causal, ya que a mayor cantidad de pesos, siempre vamos a poder convertir una mayor cantidad de Dólares
* Se observa una correlación intermedia o moderada entre el "Cantidad de Celulares" de las "Entregas" y el "Peso Bruto", "Valor Total en Pesos" y "Valor Total en Dolares" que hay en cada una de ellas
* Las correlaciones más débiles se da entre "Días Primera Visita" y "Puntualidad Primera Visita". Eso es totalmente lógico, porque mientras más días se tarda en visitar un cliente, menos probabilidad existe de que la puntualidad se cumpla
* También existe una correlación muy débil entre la "Cantidad de Celulares" y la "Cantidad de Accesorios"

***Ingeniería de Atributos:***

* Se utilizó One Hot Encoder para los datos categóricos con pocas clasificaciones
* Se utilizó Label Encoder para los datos categóricos con muchas clasificaciones
* Se utilizó Ordinal Encoding para un campo jerárquico

***Entrenamiento y Testeo:***

Se han entrenado, testeado y medidos los siguientes modelos de clasificación:

* Árbol de decisiones
* Random Forest
* KNN
* Regresión Logística

Estas son las métricas obtenidas:

Captura de pantalla de un videojuego

Descripción generada automáticamente

En base a las métricas obtenidas, y también teniendo en cuenta el consumo de Recursos al momento de las distintas ejecuciones, el **ARBOL DE DECISIÓN** es el modelo con mejores prestaciones. La profundidad sugerida (5) a través de la optimización, fué clave para que el mismo sea muy rápido y sencillo de interpretar respecto del resto.

Todos los métodos tienen métricas muy similares, pero llama la atención la **"Regresión Logística"** ya que no hay en la matriz de confusión "Verdaderos Positivos" ni "Falsos Negativos"