**Preguntas de Programación :**

**#Pregunta 1**

Código en Python

def parse(word):  
 size = len(word)  
 if size <= 10:  
 return word  
  
 return '{}{}{}'.format(word[0], size - 2, word[-1])  
  
  
print(parse("localización"))  
print(parse("internacionalización"))

**#Pregunta 2**

Código en Python

import math  
  
  
def search\_8(num, veces=0, floors=8888888888):  
 if '8' in str(num) or math.fabs(num) > floors:  
 return veces  
  
 return search\_8(num + 1, veces + 1)  
  
  
print(search\_8(0))  
print(search\_8(179))  
print(search\_8(10e9))  
print(search\_8(666667))  
print(search\_8(-999996))

**Preguntas Generales:**

#Pregunta 1

**¿Qué es Overfitting y con qué métodos detectan este problema?**

En el estudio de los algoritmos supervisados o modelos de clasificación nos podemos encontrar con dos terminologías overfitting y underfitting, el overfitting es cuando un modelo de clasificación ha sobreentrenado o sobre ajustado (casi memorizando los elementos del train) y cuando los probamos sobre nuevos registros o de validación el poder de clasificación de deteriora considerablemente, comparando indicadores del entrenamiento con indicadores de validación. Cuando ocurre este problema se dice que un modelo de clasificación es no generalizable o extrapolable puesto que nuestros resultados no serían estables.

**Formas para evitar overfiting**

* Cantidad de Variables: Es importante contar con un número apropiado de variables para las pruebas de training y testing. En algunos casos en conveniente reducir la cantidad de features. Se puede Emplear Herramientas de reducción de dimensionalidad como ACP y AF.
* Equilibrio entre Sesgo y Varianza: Se debe de tener un equilibrio al momento de elaborar un modelo hay que tener en cuenta la complejidad del mismo y la capacidad para generalizar su aplicación en data distinta.
* Conjunto de Validación de Datos: Consiste en subdividir el conjunto de datos en una muestra para training y otra para test, lo cual permitirá obtener una valoración de aciertos/fallos real del modelo y también nos permitirá detectar fácilmente efectos del overfitting /underfitting.
* Si los efectos del overfitting /underfitting.aún continúan, despues de realizar una partición del dataset entre el train y test (siguiendo un porcentaje de 70 - 30 / 80-20 según se decida), se recomienda particionar cada periodo del dataset respetando esta partición.

**Métodos de deteccón**

* Cross -Validation: Realizar
* Partición Muestral:
* Testear sobre submuestras fuera del período de entrenamiento
* Indicadores del Train son demasiado distantes y buenos respecto al Test o validación.

#Pregunta 2

**¿Qué metodologías de modelamiento machine learning conoces?**

Entre las metodologías de machine learning y en general para cualquier proyecto de minería de datos podemos resaltar a CRISP DM y SEMMA que son las más conocidas, estas metodologías se basan principalmente en el proceso de extracción de conocimiento (KDD) que pasaremos a describir a continuación:

Consta de seis fases o etapas que si bien las puedes desarrollar secuencialmente y en orden, esto no impide que puedas retroceder a fases previas en busca de mejores resultados. Las seis fases son:

**Comprensión del negocio:** En esta fase se definen tanto los objetivos del negocio como los objetivos del proyecto de minería de datos, asimismo se hace una evaluación de los recursos con los que se cuenta para definir el plan del proyecto.

**Comprensión de los datos:** En esta etapa lo que se pretende es tenner un conocimiento de la data con la que se cuenta, entenderla, describirla y comprender la utilidad que podría tener para alcanzar los objetivos del negocio)

**Preparación de los datos:** Es la etapa que más tiempo consume con el fin de quedarse con un set de datos con valor agregado donde se aplicara el modelo, los pasos que se siguen son:

* Selección de los datos
* Limpieza de datos
* Construcción de datos
* Integración de datos
* Formateo de datos

**Modelado**: En esta etapa se aplica los diferentes algoritmos de modelamiento al dataset que se obtuvo de la fase previa.

**Evaluación:** Se evalúan los modelos de la fase anterior para determinar si son útiles a las necesidades del negocio.

**Despliegue**: se implementa el modelo seleccionado, integrándose en las tareas de toma de decisiones de la organización.

#Pregunta 3

**¿Qué algoritmos de machine learning conocen? ¿Qué algoritmos de machine learning han usado? Explique brevemente uno de ellos.**

Los Algoritmos de machine learning que conocemos y que hemos aplicado ya sea en Python o R son los siguientes:

Algoritmos de Regresión:

* Regresión Lineal
* Regresión Logística

Algoritmos de Árboles de Decisión

* Árboles de clasificación y Regresión (CART)
* Random Forest
* Algoritmo de árbol de Clasificación Boosting

XGBoost, GBM, AdaBoost

Algoritmos Bayesianos

* Naive Bayes

Algoritmos de Clustering

* K-means

K-medians

Algoritmo de Redes Neuronales

* Backpropagation

Algoritmos de Reducción de Dimensión

* Análisis de componentes principales
* Análisis Factorial

El algoritmo que se va explicar es el **Gradient Boosted Decision Trees** el cual es una técnica de aprendizaje automático para problemas de regresión o clasificación, que produce un modelo de predicción en forma de un conjunto de modelos de predicción débiles. El GBDT utiliza árboles de decisión como modelos predictivos débiles o “weak learners”en el aumento de gradiente, y es uno de los algoritmos de aprendizaje más utilizados en el aprendizaje automático en la actualidad.

La idea general del método es el *entrenamiento aditivo* . En cada iteración, un nuevo árbol aprende los gradientes de los residuos entre los valores objetivo y los valores actuales predichos, y luego el algoritmo realiza el descenso del gradiente en función de los gradientes aprendidos, es decir existe una doble ponderación , la primera ponderación es de observaciones mal clasificadas y la segunda ponderación es respecto a los clasificadores. Cada clasificador recibe una ponderación o puntuación de acuerdo a cuánto se haya o no equivocado y así se toma la decisión final.

**Algoritmos de Machine Learning que usan Variables Cualitativas:**

* -Regresión Logística
* -Naive Bayes
* -Árboles de clasificación y Regresión (CART)
* -Arbol de Decisión CHAID
* -Naive Bayes
* -Random Forest

#Pregunta 4

**De los nombrados en la pregunta anterior, ¿Cuáles se usan para variables cualitativas?**

El Modelo de Regresión Logística, es uno de los modelos que acepta variables cualitativas tanto en la variable dependiente como en las covariables o variables independientes. Generalizando la gran mayoría de algoritmos utilizan o pueden trabajar con variables cualitativas en el target o variable dependiente o en las covariables, eso sí existen algoritmos tipo al árbol Chaid que al basarse en pruebas Chi - cuadrado son muy robustos o trabajan mejor con variables cualitativas.

#Pregunta 5

**Existen entornos en la nube para desarrollar modelos Machine Learning. Describa de brevemente uno de ellos.**

Existen diversas formas de desarrollar modelos de Machine Learning en la nube, hay empresas que brindan el servicio, algunos de estos son Amazon Web Service, Microsoft Azure y Google Cloud. Cada uno de ellos tienes ya incorporados algoritmos de Machine Learning. A continuación se va a describir el de Amazon Web Service.

**AWS** tiene entornos de pago para poder ejecutar modelos de machine learning, también tiene modelos incorporados, los cuales se pueden utilizar en forma de pago. Otra ventaja es que se puede comprar capacidad de procesamiento en la nube, si en caso quisieras subir código para ejecutar.

**Kaggle:** Es otra herramienta para correr código en la nube de forma gratuita, ya que nos da espacio de procesamiento y almacenamiento. Se puede correr en su notebook código en python.

#Pregunta 6:

**La seguridad de los datos es muy importante cuando se habla de Cloud. ¿Qué opinas sobre el**

**RGPD (Reglamento General de Protección de Datos)? Responda brevemente.**

Para empezar dejar bien en claro que en este contexto un dato personal es toda aquella data obtenida directa o indirectamente que puede ser usada para identificarse.

Este reglamento entra en vigor en el instante en que cualquier entidad inicie una interacción con los datos personales de los ciudadanos de países europeos designado. Donde la finalidad principal se centra en que cada ciudadano pueda gestionar e inspeccionar de manera sencilla la forma en

que se tratan su información personal. Y claro, esto ha causado todo un revuelo en los flujos de trabajo de estas entidades, intimidadas

por las altas sanciones económicas en vigencia. Un ejemplo claro en mi centro de Labores ,Redacción del Grupo Comercio, donde se tuvo que armar un proyecto de alta prioridad para lograr mostrar una ventana emergente(solo activa en países designados) que pida consentimiento del

uso de cookies que indirectamente por metadata interactuaba con ciertos datos personales.

Asimismo, Con respecto al Reglamento General de Protección de datos y la Nube, la posición que se toma es que la Nube ayuda al cumplimiento del RGPD ya que contribuye de forma más sencilla de catalogar y localizar los datos existentes en una empresa, de esta forma garantizamos la seguridad y privacidad. Un ejemplo de ello son las 4 fases propuestas por Microsoft, que tienen como objetivo que las empresas puedan adoptar de manera rápida las tecnologías y procesos que garanticen el cumplimiento de GDPR basados en soluciones en Cloud.

• Descubrir: el objetivo de esta fase es identificar qué datos personales tiene la empresa, crear un inventario, así como identificar los sistemas donde se recopilan y almacenan.

• Administrar: en esta fase, las organizaciones definen la gobernanza de los datos y su clasificación, así como la definición de políticas, roles y responsabilidades.

• Proteger: las empresas y entidades públicas deben establecer mecanismos para prevenir, detectar y responder a vulnerabilidades y violaciones de la información.

• Informar: las organizaciones deberán mostrar mayor transparencia en la forma en que controla los datos personales, así como garantizar que mantiene documentación completa sobre sus procesos y el uso de sus datos personales.

#Pregunta 7

**Si fueras parte de Movistar. ¿Qué aportes implementarías en la empresa?**

Si fuéramos parte de movistar, lo primero que haríamos son reuniones con las diferentes áreas de la empresa, con el objetivo de recoger los problemas que tienen a nivel de información y saber con qué recursos contamos para la propuesta en soluciones innovadoras.

Tenemos cinco claras propuestas para movistar:

**Primera Propuesta: Predecir la Deserción de Clientes**

Nuestra Primera sugerencia es implementar modelos de machine learning para predecir la deserción de clientes, ya sea cambio de plan, cambio a otro operador; haciendo un análisis de su comportamiento con data histórica.

Con el Objetivo de Dirigir campañas a los clientes que tengan mayor probabilidad a irse de la compañía.

**Segunda Propuesta: Sistema de Recomendación**

Realizar un sistema de recomendación sobre los modelos de equipos con los cuales cuenta movistar y el perfil de usuario que se tiene. De esta forma se va a realizar las llamadas promocionales a los usuarios, pero indicando que se tiene el equipo ideal, que se ajusta a sus necesidades, de esa forma logramos que las llamadas a los usuarios sea más personalizadas.

**Tercera Propuesta: Determinar líderes de Comunidades con SNA**

Podríamos mediante Social Network Analysis determinar primero comunidades o grupos de personas por las interrelaciones que existen entre ellas respecto a cómo se comunican mediante redes sociales y así determinar perfiles de líderes, seguidores y demás. Esto cruzado con los modelos de Churn podría darme muy buenas estrategias de fidelización o retención.

**Cuarta Propuesta: Modelos de Cross Selling de Adicionales**

Mediante análisis de asociación o market basquet analysis combinados con modelos predictivos podríamos saber cuáles son los siguientes servicios o adicionales que el cliente podrá adquirir , asu vez podríamos con ésto ofrecerle ciertas aplicaciones propias que nos permita elevar la facturación o ticket promedio.

**Quinta Propuesta: Análisis de Sentimientos en las llamadas o reclamos**

A través de ésto podríamos saber qué factores o características influyen en la fuga de los clientes o en la cancelación de sus productos y así con ésta información también ponerla de input en los modelos predictivos o combinar tanto información estructurada como no estructurada.