Implementación de Modelos de Clasificación para la Predicción de Riesgos Crediticios

1st Lucas Ramos, Jhamir
Facultad de Ciencias e Ingeniería
Pontificia Universidad Católica del Perú
San Miguel, Lima
arturo.lucas@pucp.edu.pe

2nd Reyes Burga, Andrea
Facultad de Ciencias e Ingeniería
Pontificia Universidad Católica del Perú
San Miguel, Lima
andrea.reyes@pucp.edu.pe

3rd Espinoza Concha, Kaytlin
Facultad de Ciencias e Ingeniería
Pontificia Universidad Católica del Perú
San Miguel, Lima
kaytlin.espinoza@pucp.edu.pe

4th Guzmán Tito, Jhair
Facultad de Ciencias e Ingeniería
Pontificia Universidad Católica del Perú
San Miguel, Lima
jhair.guzmant@pucp.edu.pe

I. Introducción

En el contexto de créditos bancarios es importante que el prestamista (banco) evalúe las solicitudes de préstamo según el perfil del prestatario (solicitante) para así tomar una mejor decisión acerca de la aprobación del crédito o no.

La evaluación del perfil del prestatario puede realizarse tomando en consideración datos como la edad, situación económica, propósito del crédito, etc, para que finalmente cada persona sea clasificada de acuerdo a un tipo de riesgo los cuales pueden ser :

- No riesgosa (good risk) : mayor probabilidad de que el prestatario si pague el préstamo, lo que se traduce en ganancias para el banco.
- Riesgosa (bad risk) : mayor probabilidad de que el prestatario no pague el préstamo, lo que se traduce en pérdidas para el banco.

El presente informe se enfoca en el análisis y uso de técnicas aplicadas para el diseño, implementación y evaluación de algoritmos de aprendizaje de máquina haciendo uso del conjunto de datos 'German Credit Risk' con el objetivo de predecir la existencia de un riesgo crediticio de acuerdo al perfil de un conjunto de prestatarios.

El objetivo principal de este informe es aplicar y comparar distintas técnicas de Machine Learning como modelos de clasificación en una tarea de clasificación sobre un conjunto de datos brindado. En base a los modelos implementados, se realizará una comparación entre los resultados de cada uno de estos con el objetivo de concluir cuál de ellos fue el que obtuvo un mayor desempeño. Por otro lado, se obtendrán los atributos más influyentes del conjunto de datos que determinen el resultado de la asignación de crédito lo que ayudará a obtener conclusiones acerca de cómo influyen los parámetros en la optimización y comportamiento de los modelos de clasificación.

II. Estado del arte

Los artículos revisados utilizaron el mismo conjunto de datos para el problema por lo que fueron útiles como una guía de cómo se estructura un proyecto de aplicación de modelos de clasificación.

A. Arículo 1: Maintaining the Integrity of the Specifications

Análisis de las características

La primera fase del proyecto es realizar una revisión de los tipos de datos que se tienen, la cantidad de nulos y cantidad de clases que hay por cada característica. En la segunda fase se pasó a una revisión detallada por cada característica, obteniendo la cantidad de datos clasificados como crédito riesgoso y no riesgoso para cada una de sus clases, y comparando la cantidad solicitada de crédito por cada clase según si se había dividido en crédito riesgoso o no riesgoso. Luego también se realizan otro tipos de comparaciones combinando las características como edad con cuenta bancaria, tipo de vivienda con tipo de trabajo, entre otras.

Modelos empleados

Primero realiza una comparación entre varias modelos, entre los cuales están el KNN, el Decision Tree, Random Forest, Logistic Regression. Compara estos modelos empleando sus características base y realiza un boxplot donde obtiene que el mejor modelo de los mencionados anteriormente sería decision tree. Segundo decide cambiar características como max_depth, n_estimators, ma_features en el modelo de Random Forest para mejorar el resultado.

Aporte

El aporte obtenido de este artículo es la forma de evaluar las características para entender cómo influyen cada una de ellas en la clasificación del crédito como riesgoso o no riesgoso.

B. Artículo 2: German Credit Analysis, A Risk Perspective

Descripción de las características

La primera fase del proyecto es ver los datos, que variables son numéricas o categóricas y que columnas tienen valores nulos. En la segunda fase realizó un análisis por tipo de grupo: análisis por sexo, análisis por edad (crea grupo categóricos) y análisis de riqueza (analiza las cuentas corrientes) para poder ver cómo estas características contribuyen al riesgo de los préstamos otorgados a los clientes. En la tercera fase analiza los préstamos de alto y bajo riesgo para poder encontrar patrones que puedan describir algún tipo de correlación con los valores de salida. También, en esta fase, explora los propósitos de los préstamos para ver qué propósitos son los que tienen más probabilidades de generar mayor riesgo En la cuarta y última fase realiza su modelo predictivo.

Aporte

El aporte obtenido de este artículo es el análisis de las características, nos ayudó a tener una mejor noción de las características que podrían ser las más influyentes.

III. Diseño del experimento

A. Descripción de los atributos

La información del conjunto de datos pertenece a personas que pretenden adquirir un crédito bancario y son clasificados según el riesgo como bueno o malo. Los atributos encontrados en el archivo son los siguientes:

- Edad: Está en el rango de 19-75 años.
- Sexo: Puede ser femenino o masculino.
- Trabajo: Es el tipo de trabajo que realiza la persona.
  - 0: La persona realiza un trabajo no cualificado y no es residente.
  - 1: La persona realiza un trabajo no cualificado y es residente.
  - 2: La persona realiza un trabajo cualificado.
  - 3: La persona realiza un trabajo altamente cualificado.
- Alojamiento: Puede ser propio, rentado o gratis (programa del gobierno).
- Cuenta de ahorro: Puede ser un monto pequeño, moderado,grande, muy grande.
- Cuenta corriente: Puede ser un monto pequeño, moderado y grande.
- Cantidad de crédito: Está en la moneda del país (DM - Deutsch Mark).
- Duración: Está registrado en meses.
- Razón: Puede ser por carros, muebles/equipos, radio/televisión, reparos, - usos domésticos, educación,negocios, vacaciones/otros.
- Riesgo: Determina si el crédito prestado es de riego o no. Puede ser bueno o malo.

Clasificación de los atributos

De acuerdo a la información brindada se obtuvo que los datos se organizan de la siguiente manera según su tipo:

- Cardinal: Edad, Cantidad de crédito, Duración.
- Binario: Sexo, Riesgo.
- Nominal: Trabajo, Alojamiento, Cuenta de ahorro, Cuenta corriente, Razón.

Cantidad de datos en cada atributo

De acuerdo a la información del conjunto de datos se obtuvieron los siguientes resultados con respecto a cada atributo.

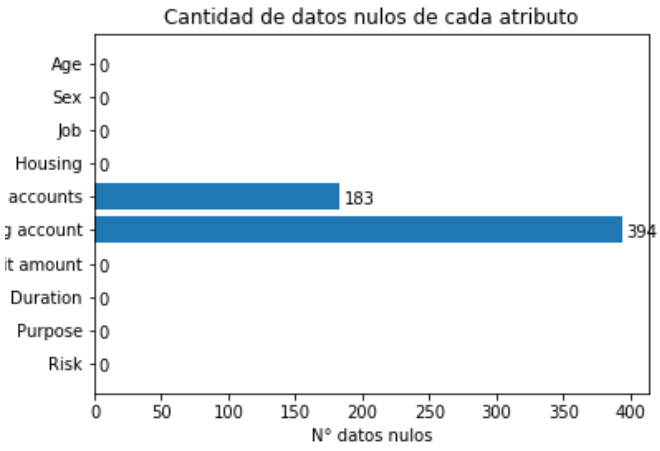
- Edad: 1000
- Sexo: 1000
- Trabajo: 1000
- Alojamiento: 1000
- Cuenta de ahorro: 817
- Cuenta corriente: 606
- Cantidad de crédito: 1000
- Duración: 1000
- Razón: 1000
- Riesgo: 1000

B. Metodología

Estrategia para datos faltantes

Mediante la evaluación del conjunto de datos se ha notado que existen atributos que tienen información faltante (Cuenta de ahorro, Cuenta corriente), estos datos faltantes podrían llevar a nuestros modelos a inconsistencias y clasificaciones erróneas por lo cual se ha optado por utilizar una técnica de administración de datos faltantes descrita el en articulo Handling Missing Values in Machine Learning (Georgios Drakos) que consiste en llenar los datos faltantes con la moda de la clase, en este caso las modas de cada clase con datos faltantes son las siguientes:

- Cuenta de ahorro: Little
- Cuenta corriente: Little



Procesamiento de la Información

De los 9 atributos mencionados anteriormente, se decidió eliminar el atributo Sexo pues no se consideraba importante

en la evaluación del riesgo y podía influir en la decisión de manera errónea. Luego, se convirtieron los datos cualitativos como Alojamiento, Cuenta de ahorro, Cuenta corriente,Razón, Riesgo a cuantitativos mediante la tde label encoder para que la clasificación anterior en cada atributo est representada mediante números.

Separación de la información en entrenamiento y validación

Con el objetivo de evitar problemas como el overfitting y poder estimar con mayor veracidad la correctitud de nuestros modelos de clasificación, se utilizó un método de splitting que consiste en la división del conjunto de datos.

Primero dividimos toda la información en conjuntos llamados X e Y donde X tendrá los atributos del 1-8 y Y tendrá el atributo target que sería Riesgo. Luego, con el objetivo de entrenar y luego validar nuestros modelos de clasificación, se realiza la separación de X e Y en dos subconjuntos, uno de ellos servirá para el entrenamiento de los modelos de clasificación y el otro para la validación de los resultados, estos subconjuntos se describen a continuación:

- Tamaño de la información de entrenamiento: 700 x 8
- Tamaño de la información de validación: 300 x 8

Métodos empleados

como se mencionó anteriormente, para la realización de la tarea de clasificación se utilizó principalmente tres métodos Logistic Regression, Random Forest y KNN los cuales se describen a continuación.

Logistic Regression

Se usará este modelo porque es un clasificador para problemas donde el target es de clase binaria y las variables de entrada tienen diferentes clases. A continuación se muestran algunos de los parámetros más relevantes para este modelo de clasificación.

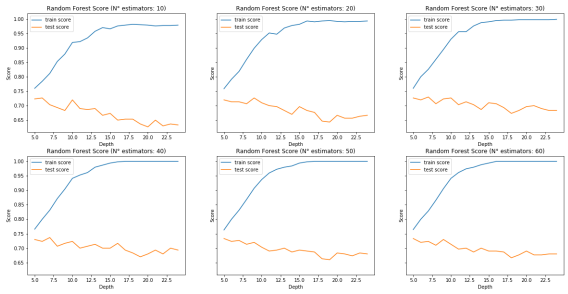
random_state = 7 que es la semilla del generador de números pseudoaleatorios para usar cuando se barajan los datos. solver = ”liblinear” que es el algoritmo a utilizar en el problema de optimización y liblinear es una buena opción para pequeños conjuntos de datos.

Random Forest

Se usará este modelo porque utiliza múltiples Decision Trees, previene el overfitting y sirve para problemas de clasificación. A continuación se muestran algunos de los parámetros más relevantes para este modelo de clasificación.

criterion = gini nos ayuda a saber cuán diversos están los datos. max_depth = 10 que fue un valor elegido arbitrariamente. n_estimators = 50 que también fue un valor elegido arbitrariamente. n_jobs = -1 random7_state = 7

Después se demostró que los valores elegidos para los argumentos max_depth y n_estimators fueron los correctos para el Random Forest, ya que se puede apreciar que el score de la data de entrenamiento se aproxima a 1 a partir de una profundidad de 15.



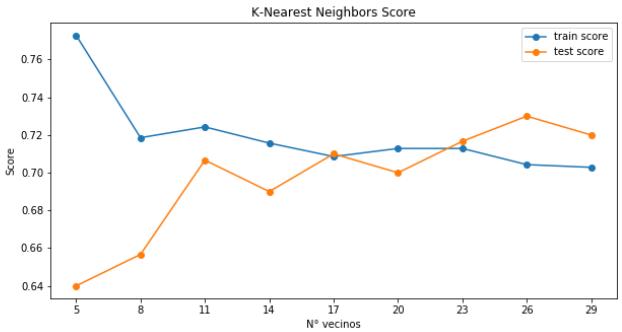
Clasificación KNN

Se usará este modelo porque es simple de implementar, es intuitivo, clasifica la nueva información según los registros que conoce, se adapta según la nueva información que va obteniendo y sirve para problemas de clasificación. A continuación se muestran algunos de los parámetros más relevantes para este modelo de clasificación.

n_neighbors = 17

p = 3

El primer parámetro se escogió al comparar la cantidad de vecinos con la que se obtenía el mejor score sin llegar al overfitting. El mejor score obtenido fue con 17 vecinos.

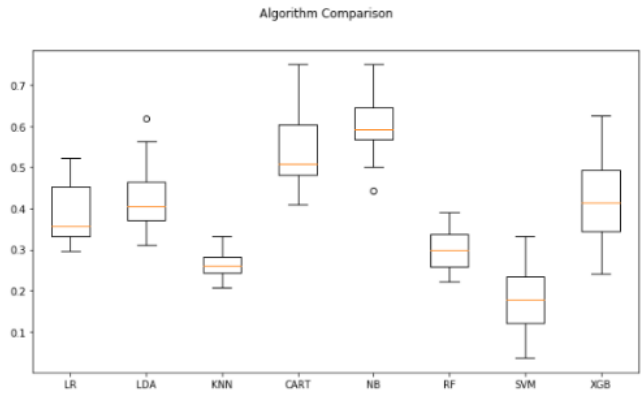


El segundo parámetro se refiere a la ecuación para hallar la distancia, siendo en este caso la distancia de Minkowski.

C. Experimentación y resultados

Reproducción de resultados reportados en un artículo científico anterior

El artículo Predicting Credit Risk escrito por Leonardo Ferreira realiza una comparación entre varios modelos de clasificación como son KNN, Random Forest, Logistic Regression, GaussianNB, etc, pero ahora nos centraremos en estas tres primeras.



Se puede notar que de acuerdo a los resultados obtenidos por dicho autor, Logistic Regression es el clasificador que obtiene los mejores resultados pero estos aun asi no obtienen resultados óptimos, pues obtienen una exactitud por debajo del 50%, esto a pesar de que el autor realiza el método de KFold.

Resultados de nuestro modelo - Comparación entre los modelos empleados

Logistic Regression:
Logistic Regression train SCORE: 0.71
Logistic Regression test SCORE: 0.71

Random Forest:
Random Forest Classifier train SCORE: 0.94
Random Forest Classifier test SCORE: 0.70

KNN:
K-Nearest Neighbors train SCORE: 0.7085714285714285
K-Nearest Neighbors test SCORE: 0.71

Después de evaluar los modelos, se realiza una comparación del score de los tres modelos obteniendo un mejor resultado con el modelo de KNN; pero como solo se ha probado con una forma de partición de la data, se realiza Cross Validation, Matriz de confusión y AUC ROC para obtener un resultado más completo.

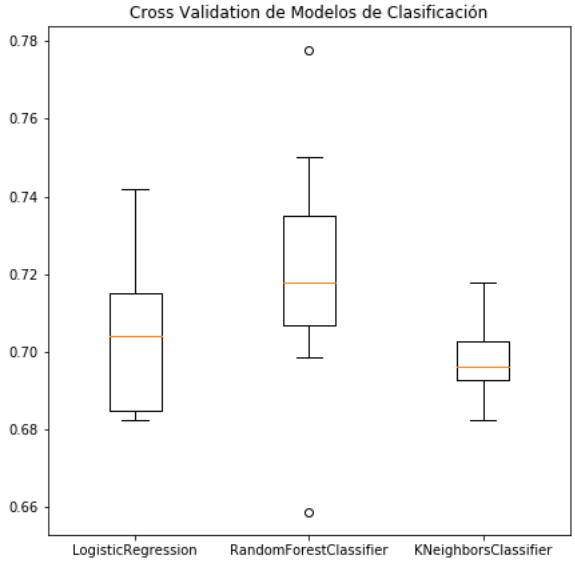
Cross Validation (K-FOLDS)

En este caso, se usará un split de 8-folds para obtener un número considerable en las evaluaciones que realizará cada modelo y para obtener una mejor descripción de este tipo de evaluación se realizó un Boxplot.

Logistic Regression:
Average CV score LogisticRegression : 0.704

Random Forest:
Average CV score RandomForestClassifier: 0.720

KNN:
Average CV score KNeighborsClassifier : 0.698



Como se aprecia en la imagen, el modelo Random Forest Classifier ahora es el de mayor score, puesto que antes se indicaba al KNN como el mejor.

Matriz de confusión

En este caso, se busca obtener el modelo que realiza una mejor clasificación para la clase Bad Risk en riesgo, ya que se presentan menos registros clasificados con dicha clase en la data.

Logistic Regression

	Precisión	Recall
Bad Risk	0.55	0.12
Good Risk	0.72	0.96

Random Forest

	Precisión	Recall
Bad Risk	0.51	0.21
Good Risk	0.73	0.91

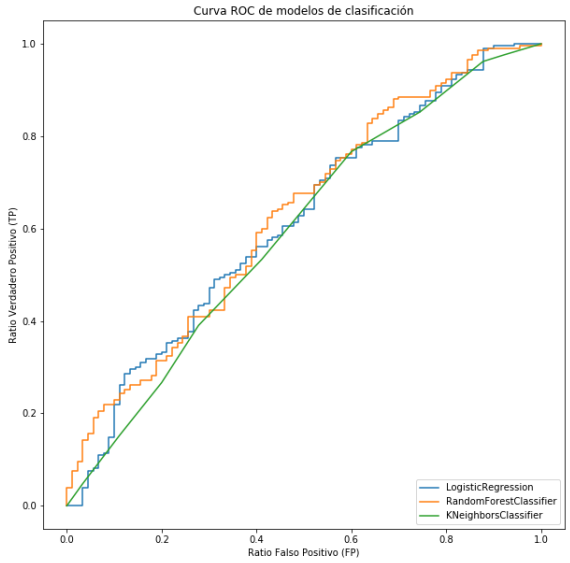
K-Nearest Neighbors

	Precisión	Recall
Bad Risk	0.58	0.12
Good Risk	0.72	0.96

Al observar estas tres imágenes, los tres modelos no logran clasificar correctamente a los que pertenecen a la categoría de Bad Risk.

Curva ROC

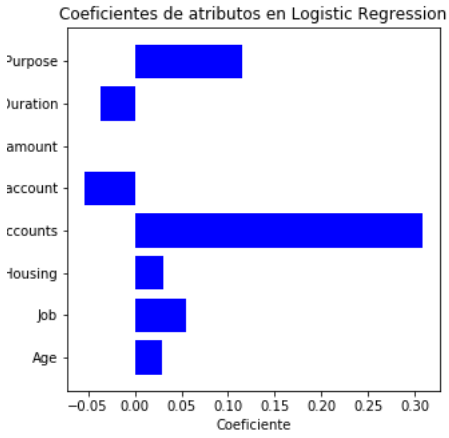
Luego se realizó una Curva Roc donde se compara a los tres modelos.



Comparando las AUC de cada modelo de clasificación, RandomForest es quien clasifica mejor a la data entregada, y KNeighbors es el modelo que menor lo realiza correctamente.

Atributos más influyentes

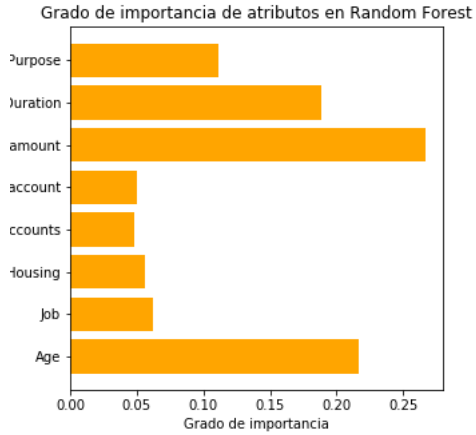
Logistic Regression



Como se puede apreciar en el gráfico, los atributos más influyentes son Saving accounts y Purpose.

Random Forest

Este modelo otorga la facilidad de obtener los grados de importancia de cada atributo de la data con la que ha sido entrenado.



Como se puede observar en el gráfico, los atributos más influyentes son Credit amount, Age, Duration y Purpose.

K-Nearest Neighbors

En este caso, no se puede determinar qué atributos son los más influyentes debido a que este modelo es un clasificador ”perezoso”, no construye el modelo explícitamente, por lo que se optará por analizar a este modelo con los atributos influyentes comunes en los dos modelos anteriores.

Comparación de línea base y resultados propios

En la línea base utilizada que es el artículo escrito por Leonardo Ferreira: El modelo con mejores resultados es el Logistic Regression, después el Random Forest y luego está el KNN. En cambio,en los resultados propios el modelo con mejores resultados es el Random Forest y después el Logistic Regression. Lo que resulta igual es que entre los tres modelos el que tiene peores resultados es el KNN.

D. Discusión

Interpretación de los resultado obtenidos

Al realizar una primera evaluación, el modelo con que se obtenía un mejor resultado era KNN. El problema es que cuando se evalúa con una sola forma de división no se puede optar como mejor ese clasificador, porque el mejor clasificador debe ser independiente de la partición de los datos en entrenamiento y prueba.

Los resultados obtenidos en Cross Validation son obtenidos al evaluar los modelos con diferentes particiones de la data. De esto se obtuvo como mejor modelo al Random Forest Classifier y hubo una diferencia un poco más grande con el score del Logistic Regression. Esto es posible pues al obtener varios Decision Tree que utilizan diferentes características de la data se evalúa cuáles de ellas ayudan a obtener un mejor resultado y se les da mayor importancia. Así mismo, al utilizarlos con diferentes grupos de entrenamiento, se logra obtener un más alto score pues se realiza un estudio más variado donde con ciertas particiones se lograrían mejores resultados.

Los resultados obtenidos con la matriz de confusión ayudan a evaluar con qué tipo de datos se realizan más errores, es

decir, con los datos clasificados como Good Risk o como Bad Risk. En este caso, en los tres modelos se obtienen más errores con Bad Risk por el hecho que hay menos datos en esta clasificación, por lo que no se saben muchas características que evaluarían a un nuevo dato como Bad Risk. Por otro lado, los tres modelos logran clasificar de manera correcta a la mayoría de datos clasificados como Good Risk pues hay más datos de este grupo en el grupo de entrenamiento.

Los resultados de la curva roc realizada muestran que Random Forest clasifica mejor a la data entregada, y KNN es el modelo que mejor lo realiza correctamente. Por lo que el clasificador con mayor sensibilidad sería el Random Forest.

En conclusión el Random Forest sería el mejor clasificador para este problema pues es el que mejores resultados ha obtenido en los métodos usados.

Identificación y visualización de ejemplos en los que tienen dificultad los modelos ensayados. A qué se podría atribuir?

Como ya se describió con anterioridad existen datos con valor nulo (account), los cuales tienen un impacto notable en los resultados de nuestros clasificadores, ya que estos tienen una importancia mayor en algunos algoritmos como Random Forest y Knn , por lo que al rellenar estos espacios con datos como la media o moda puede ocurrir ciertas imprecisiones en el resultado de los clasificadores.

Por otro lado también cabe resaltar que de acuerdo a las estadísticas del conjunto de datos se puede notar que existe una cantidad mucho mayor de Good Risk en comparación a los Bad Risk, lo cual produce que en general los algoritmos clasifiquen de mejor manera a los datos que pertenecen a la clasificación de Good Risk, mientras que para Bad Risk existe un rendimiento menor.

E. Conclusiones y trabajos futuros

Las primeras conclusiones obtenidas fueron acerca de la data observada:

Se concluye que de los 9 atributos disponibles, el atributo Sexo no se considera necesario ya que crea una partición errónea de la data, es decir que sí hablamos en el contexto de créditos bancarios, el sexo no influye en absoluto por lo cual debe ser descartado.

Se concluye que el atributo Trabajo está representado como entero pero se le considera como categórico porque representa el tipo de trabajo que realiza la persona.

F. Conclusiones para trabajos futuros

Se concluye que para trabajos futuros para realizar mejores evaluaciones de los modelos, se utilicen estrategias como Cross Validation, Curva Roc y Matriz de confusión. Así se obtendrían resultados más reales y se podría obtener el mejor modelo para el caso evaluado.

G. Referencia

Predicting Credit Risk
Leonardo Ferreira
https://www.kaggle.com/kabure/predicting-credit-risk-model-pipeline

Credit Risk Data
https://www.kaggle.com/patelfaiz007/credit-risk-data

Predicting Credit Risk Model Pipeline
Leonardo Ferreira
https://www.kaggle.com/kabure/predicting-credit-risk-model-pipeline

Credit Risk Analysis Eda
Thiago Panini
https://www.kaggle.com/thiagopanini/credit-risk-analysis-eda-full-pipeline

plot.ly
https://plot.ly/create/

Handling Missing Values In Machine Learning Part 1
Georgios Drakos
https://towardsdatascience.com/handling-missing-values-in-machine-learning-part-1-dda69d4f88ca