Detección de Transacciones Fraudulentas en Tarjetas de Crédito

Asignatura: Proyectos

Vicente Frías - Bastián Aceitón

Noviembre 2024

UNIVERSIDAD TECNICA
FEDERICO SANTA MARIA

Tabla de Contenidos

Motivación y contexto del problema

- El fraude en tarjetas de crédito es un problema grave que afecta tanto a las instituciones financieras como a los usuarios.
- Este conjunto de datos presenta transacciones que ocurrieron en dos días, donde se tienen 492 fraudes de un total de 284.807 transacciones. El conjunto de datos está altamente desbalanceado, la clase positiva (fraudes) representa el 0,172% de todas las transacciones.
- Las características V1, V2, ..., V28 son los componentes principales obtenidos con PCA, y las únicas características que no han sido transformadas son 'Time' y 'Amount'
- 'Time' contiene los segundos transcurridos entre cada transacción y la primera transacción en el conjunto de datos. La característica 'Amount' es el monto de la transacción
- La característica 'Class' es la variable de respuesta y toma el valor 1 en caso de fraude y 0 en caso contrario.

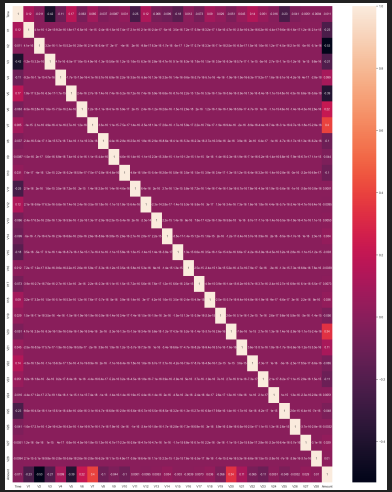
Tabla de Contenidos

2 Análisis Exploratorio De Datos

Correlación:

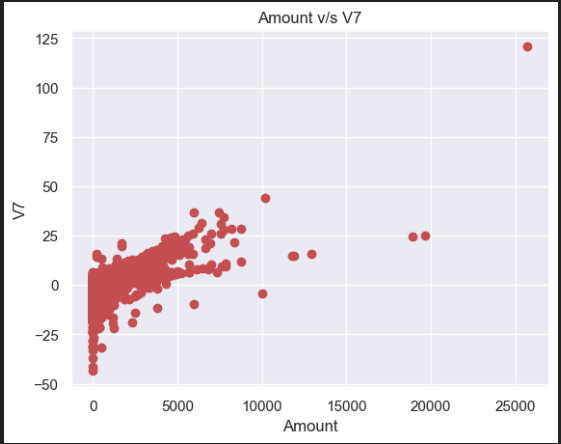
Análisis Exploratorio De Datos

2 Análisis Exploratorio De Datos

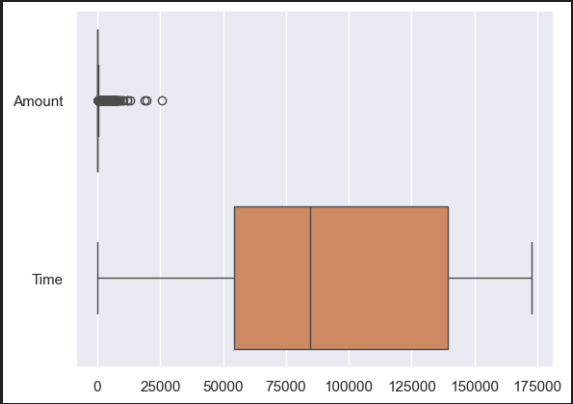


Análisis Exploratorio De Datos

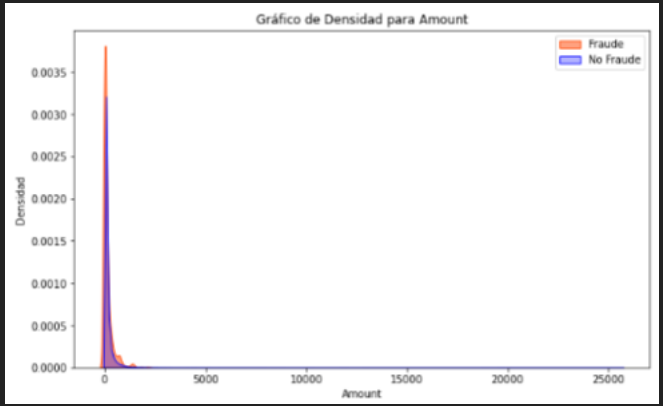
2 Análisis Exploratorio De Datos



Datos Atípicos:

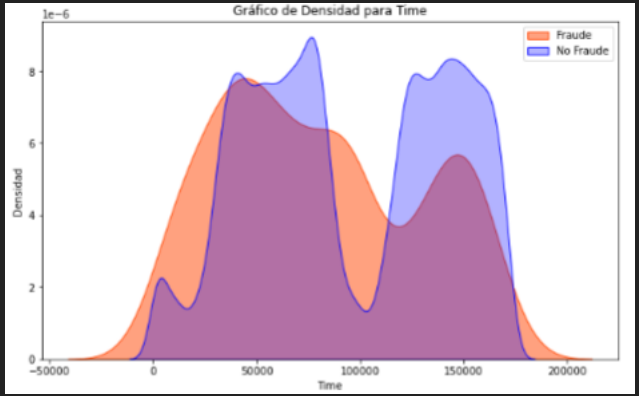


Densidades Estimadas:

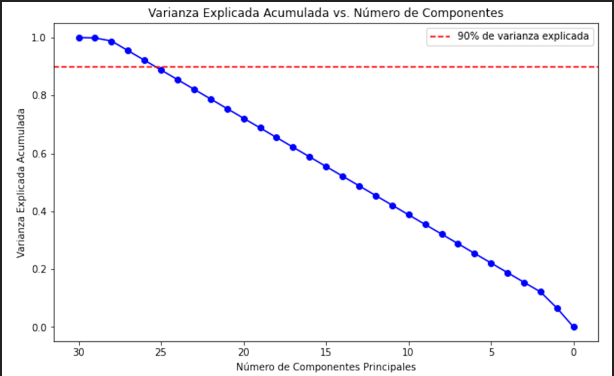


Análisis Exploratorio De Datos

2 Análisis Exploratorio De Datos



PCA:


Varianza Explicada Acumulada vs. Número de Componentes
90% de varianza explicada
Varianza Explicada Acumulada
1.0
0.8
0.6
0.4
0.2
0.0
30
25
20
15
10
5
0
Número de Componentes Principales

De donde se puede notar que el umbral de 90% se cruza con 25 componentes. Es por esto que los modelos se van a trabajar con 26 componentes, lo cual implica un 92,14% de varianza acumulada explicada, que resulta suficiente para conservar la mayoría de la información relevante y además obtener un conjunto de menos dimensionalidad.

Tabla de Contenidos

3 Métricas

- **Área bajo la curva Precision-Recall (AUPRC)**: El Área bajo la curva Precision-Recall (AUPRC) es una métrica adecuada para problemas con clases desbalanceadas. Captura el equilibrio entre precisión y recall para distintos umbrales de probabilidad.
- **F1 Score**: La puntuación F1 es la media armónica entre precisión y recall.
- **Área bajo la curva ROC (AUC-ROC)**: El AUC-ROC mide la capacidad del modelo para distinguir entre clases.
- **Exactitud (Accuracy)**: La exactitud es la proporción de predicciones correctas sobre el total de predicciones.
- **Tiempo de Clasificación**.

Tabla de Contenidos

4 Modelos

- **Modelo Naive Bayes**: El modelo de Naive Bayes es un clasificador probabilístico basado en el teorema de Bayes. Se llama 'naive' (ingenuo), porque asume que las características son independientes entre sí.
- **Gradient Boosting**: Es un método de ensamble que construye un modelo fuerte a partir de varios modelos débiles (generalmente árboles de decisión). En cada iteración, el modelo intenta corregir los errores del modelo anterior, ajustando los errores residuales.
- **Redes Neuronales**: Las Redes Neuronales son modelos de aprendizaje profundo inspirados en el cerebro humano. Están compuestas de capas de nodos (neuronas) conectadas entre sí. Cada conexión tiene un peso asociado que se ajusta en el entrenamiento.

- **Random Forest**: RandomForest es un modelo de ensamble basado en múltiples árboles de decisión, cada uno entrenado en diferentes subconjuntos aleatorios de los datos y características. La predicción final se obtiene combinando las predicciones de cada árbol mediante votación (clasificación) o promedio (regresión).
- **Árbol de Decisión (Decision Tree)**: Un Árbol de Decisión es un modelo de clasificación y regresión que divide el espacio de características en regiones homogéneas basándose en un conjunto de condiciones lógicas. Cada nodo representa una característica, y cada rama representa una decisión basada en los valores de esa característica.
- **K-Nearest Neighbors (KNN)**: Es un algoritmo de clasificación basado en la proximidad. Para clasificar una muestra, el modelo encuentra los k vecinos más cercanos en el espacio de características y asigna la clase más común entre ellos.

- **Quadratic Discriminant Analysis (QDA)**: Es un clasificador que asume que los datos siguen una distribución normal, pero a diferencia de LDA, permite que cada clase tenga su propia matriz de covarianza. El modelo clasifica los datos calculando la probabilidad de que una muestra pertenezca a una clase específica, basándose en la función discriminante cuadrática.

Tabla de Contenidos

5 Resultados

Naive Bayes:

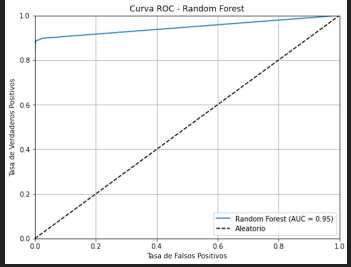
Métrica	Valor
AUPRC	0.0820
Exactitud	0.9797
Puntuación F1	0.1201
AUC	0.9568
Tiempo de entrenamiento (segundos)	0.0977
Tiempo de clasificación (segundos)	0.0289

Cuadro 1: Resultados del modelo Naive Bayes

Random Forest:

Métrica	Valor
AUPRC	0.8587
Exactitud	0.9995
Puntuación F1	0.8391
AUC	0.9474
Tiempo de entrenamiento (segundos)	178.3891
Tiempo de clasificación (segundos)	0.4508

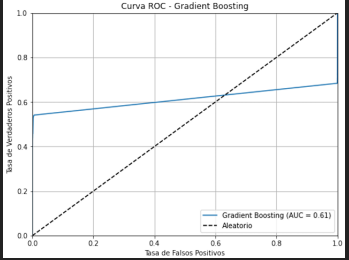
Cuadro 2: Resultados Random Forest



Gradient Boosting:

Métrica	Valor
AUPRC	0.4131
Exactitud	0.9988
Puntuación F1	0.5352
AUC	0.6118
Tiempo de entrenamiento (segundos)	234.5868
Tiempo de clasificación (segundos)	0.0529

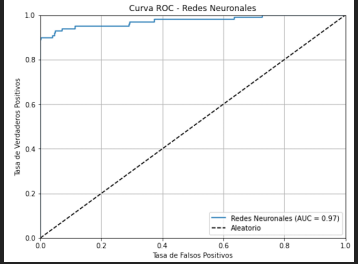
Cuadro 3: Resultados Gradient Boosting



Redes Neuronales:

Métrica	Valor
AUPRC	0.8706
Exactitud	0.9995
Puntuación F1	0.8333
AUC	0.9730
Tiempo de entrenamiento (segundos)	21.6481
Tiempo de clasificación (segundos)	0.0469

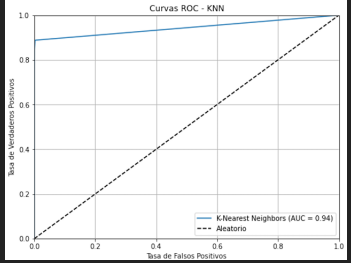
Cuadro 4: Resultados Redes Neuronales



KNN:

Métrica	Valor
AUPRC	0.8280
Exactitud	0.9994
Puntuación F1	0.7882
AUC	0.9437
Tiempo de entrenamiento (segundos)	0.0189
Tiempo de clasificación (segundos)	193.2743

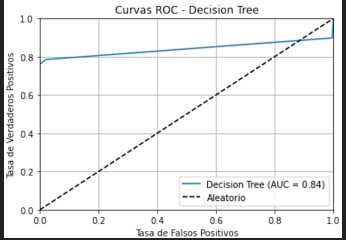
Cuadro 5: Resultados K-Nearest Neighbors (KNN)



Decision Tree:

Métrica	Valor
AUPRC	0.7094
Exactitud	0.9995
Puntuación F1	0.8249
AUC	0.8408
Tiempo de entrenamiento (segundos)	7.4451
Tiempo de clasificación (segundos)	0.0060

Cuadro 6: Resultados Decision Tree



Quadratic Discriminant Analysis:

Métrica	Valor
AUPRC	0.1419
Exactitud	0.9743
Puntuación F1	0.1074
AUC	0.9793
Tiempo de entrenamiento (segundos)	0.2922
Tiempo de clasificación (segundos)	0.0259

Cuadro 7: Resultados Quadratic Discriminant Analysis (QDA)

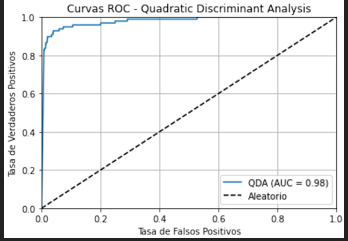


Tabla de Contenidos

6 Análisis Resultados

Análisis Resultados

- **Redes Neuronales**: Mostró ser el modelo más equilibrado en cuanto a las métricas AUPRC, Exactitud, F1 y AUC. Su rendimiento es consistentemente alto, lo que lo convierte en una opción fuerte para la detección de fraudes.
- **QDA**: Tiene el peor desempeño en términos de AUPRC, Exactitud y F1, aunque su AUC es alto. Esto sugiere que el modelo tiene dificultades para manejar clases desbalanceadas, lo que es crucial en tareas de detección de fraudes.
- **KNN**:En términos de exactitud y AUPRC muestra un buen desempeño, pero tiene un tiempo de clasificación significativamente más alto en comparación con otros modelos.
- **Decision Tree**: Muestra un rendimiento aceptable en la mayoría de las métricas, pero no destaca en ninguna de ellas de manera significativa.

- **Gradient Boosting**: Tiene un AUPRC más bajo que Redes Neuronales y KNN, lo que sugiere que no es tan efectivo en términos de precisión y recall en este caso específico.

Por lo tanto, se recomienda probar modelos como Redes Neuronales para obtener el mejor rendimiento general, aunque otros modelos como KNN o Random Forest pueden ser útiles dependiendo de los requisitos de tiempo de clasificación y eficiencia.

Tabla de Contenidos

7 Bibliografía

Bibliografía

7 Bibliografía

1. title: A first course in machine learning, author: Rogers, Simon and Girolami, Mark, year:2016, publisher:Chapman and Hall/CRC
2. title: An introduction to statistical learning, author: James, Gareth and Witten, Daniela and Hastie, Trevor and Tibshirani, Robert and others, volume: 112, year: 2013, publisher: Springer
3. title: Artificial intelligence: a modern approach, author: Russell, Stuart J and Norvig, Peter, year: 2016, publisher: Pearson

Detección de Transacciones Fraudulentas en Tarjetas de Crédito

¡Muchas Gracias!