INSTITUTO SUPERIOR POLITÉCNICO DE CÓRDOBA

TECNICATURA SUPERIOR EN CIENCIA DE DATOS E INTELIGENCIA ARTIFICIAL

Módulo de Práctica Profesionalizante

Entrega 3 - Modelado Predictivo

20/10/2024

Docentes:

- Charletti, Carlos

Integrantes:

- López, Erick
- Nüesch, Christian
- Zurita Rojo, Debora
- Galeano, Agustín

Índice

Metodología de modelado

La metodología de modelado utilizada siguió un enfoque típico y estructurado dentro de la ciencia de datos y machine learning, conocido como el ciclo de vida de un proyecto de modelado predictivo. Este proceso incluyó varias etapas que son importantes para garantizar que el modelo que se construye sea robusto, interpretable y útil para los objetivos del negocio.

Dichas etapas fueron:

- Definir el problema: Predecir el precio de los autos en base a varias características.
- Preprocesamiento de datos: Transformar variables categóricas y numéricas, limpieza de datos, división en conjuntos de entrenamiento y prueba.
- Selección de modelos: Probar diferentes algoritmos de aprendizaje (Regresión lineal múltiple, Gradient Boosting, Random Forest).
- Entrenamiento del modelo: Ajustar cada modelo utilizando los datos de entrenamiento.
- Evaluación del rendimiento: Utilizar métricas como MSE, MAE, y R^2 para comparar modelos y seleccionar el mejor.
- Optimización: Ajustar hiperparámetros (si es necesario) para mejorar el rendimiento del modelo seleccionado.
- Interpretación y conclusiones: Aplicar los resultados obtenidos al contexto del negocio.

Resultados de entrenamiento y evaluación

Se evaluó el rendimiento de cada modelo comparando las métricas obtenidas:

	MSE	MAE	R^2
Regresión lineal múltiple - Primera prueba	1424760535.80	10918.03	0.36
Regresión lineal múltiple - Tercera prueba	786232977.76	8405.88	0.65
Gradient Boosting	657591633.06	6342.08	0.71
Random Forest	956274644.38	7212.61	0.57

Selección del mejor modelo

Después de evaluar varios modelos predictivos, se optó por el que logró el mejor equilibrio entre precisión y robustez: el Gradient Boosting. Este modelo destacó por su rendimiento, alcanzando los valores más bajos de error cuadrático medio (MSE) y los más altos de coeficiente de determinación (R^2).

Si bien la tercera implementación de la regresión lineal múltiple arrojó resultados aceptables, el Gradient Boosting no solo capturó de manera más efectiva las relaciones no lineales entre las variables, sino que también mostró una menor propensión al overfitting, lo que lo convierte en la opción más sólida para la predicción de precios de autos usados.

Discusión de resultados

Lo que se hizo fue analizar los resultados de diferentes modelos de predicción aplicados a la industria automotriz, específicamente para estimar precios de autos usados. El objetivo también fue interpretar las métricas de evaluación (MSE, MAE y R^2) y su impacto en el negocio.

MSE (Error Cuadrático Medio): El modelo de Gradient Boosting obtiene el error más bajo (657 millones), lo que lo posiciona como el más preciso para predecir precios y minimizar errores, impactando favorablemente en la rentabilidad del negocio. Otros modelos, como la regresión lineal múltiple y el Random Forest, muestran un mayor margen de error.

MAE (Error Absoluto Medio): El Gradient Boosting tiene un error promedio de 6342 unidades monetarias. Este error puede ser aceptable o no, dependiendo del rango de precios de los autos. Es más preciso que los otros modelos evaluados, lo que sugiere su utilidad en la toma de decisiones.

R^2 (Coeficiente de determinación): Gradient Boosting también destaca en esta métrica con un 71% de variabilidad explicada, lo que refuerza su capacidad de capturar patrones importantes. Regresión lineal y Random Forest son menos precisos en comparación.

Para concluir, Gradient Boosting es el modelo más confiable para predicciones de precios de autos usados según los datos que poseemos, lo que permitiría ajustar precios más competitivos, optimizar ganancias y reducir errores significativos. También se sugiere ajustar hiperparámetros o mejorar la calidad de los datos si la precisión aún no es suficiente.

Enlaces

Enlace al Google Colab:
https://colab.research.google.com/drive/1v0zvU2CjvN0-HB8Y0UuBtEZNT9OJTeJ3

Enlace al repositorio en GitHub:
https://github.com/ISPC-TSCDIA/Data24_PPI

Bibliografía

Géron, A. (2020). *Aprende Machine Learning con Scikit-Learn, Keras y TensorFlow: conceptos, herramientas y técnicas para construir sistemas inteligentes.* Anaya Multimedia.

Ciencia de Datos. (2024). *Material formativo sobre estadística, algoritmos de machine learning, ciencia de datos y programación en R y Python*. https://cienciadedatos.net/

Ciencia de Datos. (2024). *Machine learning con Python y scikit-learn*. https://cienciadedatos.net/documentos/py06_machine_learning_python_scikitlearn

Ciencia de Datos. (2024). *Regresión lineal con Python*. https://cienciadedatos.net/documentos/py10-regresion-lineal-python

Zapata, J. (2023). *Clasificación en Python para Ciencia de Datos.* https://joserzapata.github.io/courses/python-ciencia-datos/clasificacion