

A Machine Learning Approach to Predicting Crude Oil Viscosity

Dylan Aygun

Abstract

Crude oil viscosity affects pipeline transport, refinery operations, and commercial valuation, but laboratory measurements can take hours or days.

Leveraging Google CoLab environment, I developed a machine learning model to predict viscosity at 40 °C using five crude oil properties: density, paraffin content, aromatic content, pour point, and vanadium concentration.

After getting data from public crude-oil assays from well-regarded data sources, I trained an XGBoost regression model on 52 light and medium crude samples.

The model achieved an R² of 0.836 on the test set. Density was the strongest predictor, and predicted viscosity increased sharply above about 0.86 g/cc. The results suggest that nonlinear modeling can provide a useful, faster complement to laboratory testing.

Introduction

Crude oil viscosity determines how easily oil flows through pipelines and affects pumping energy, blending, and refinery operations. Traditional laboratory measurements can require hours or even days.

Previous viscosity correlations often focus on narrow crude categories or assume linear relationships, even though crude composition interacts in nonlinear ways.

Research question:
Can a data-driven model estimate viscosity at 40 °C from commonly measured crude-oil properties quickly and reliably?

Methods

Data: 52 light and medium crude-oil samples compiled from publicly available assay reports.

- Inputs:
- Density at 15 °C
 - Paraffin content
 - Aromatic content
 - Pour point
 - Vanadium concentration

- Model:
- 75% training / 25% testing
 - XGBoost regression
 - Natural-log transform of viscosity
 - Hyperparameters adjusted to reduce overfitting

Evaluation:
MAE, RMSE, R², permutation feature importance, partial dependence, and SHAP analysis.

Results

Test-set performance:
R² = 0.836
MAE = 2.33 cSt
RMSE = 4.08 cSt

Density had the strongest influence on model predictions. Aromatics and vanadium followed, while paraffin content and pour point contributed less overall.

Predicted viscosity increased more sharply once density rose above about 0.86 g/cc.

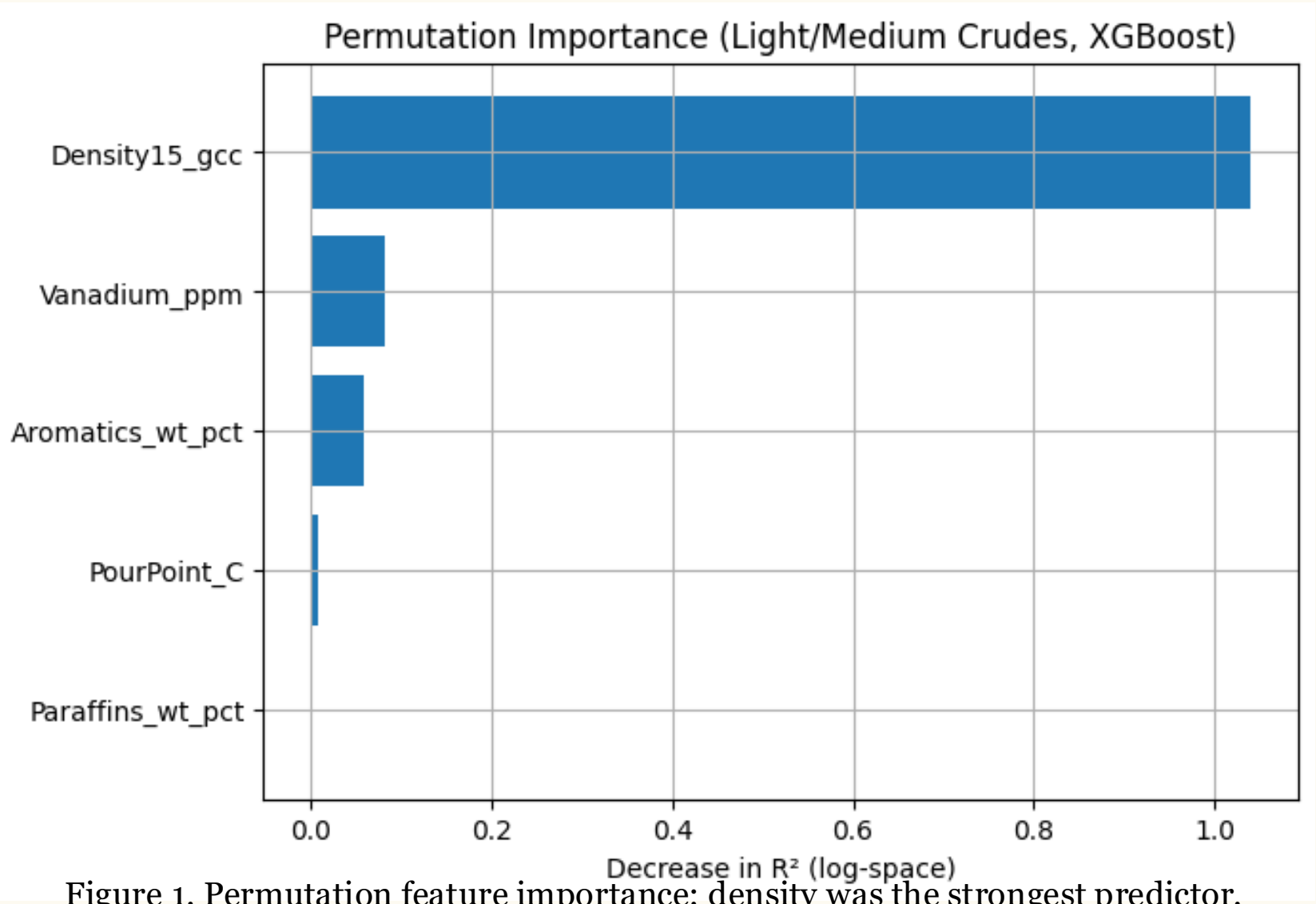


Figure 1. Permutation feature importance: density was the strongest predictor.

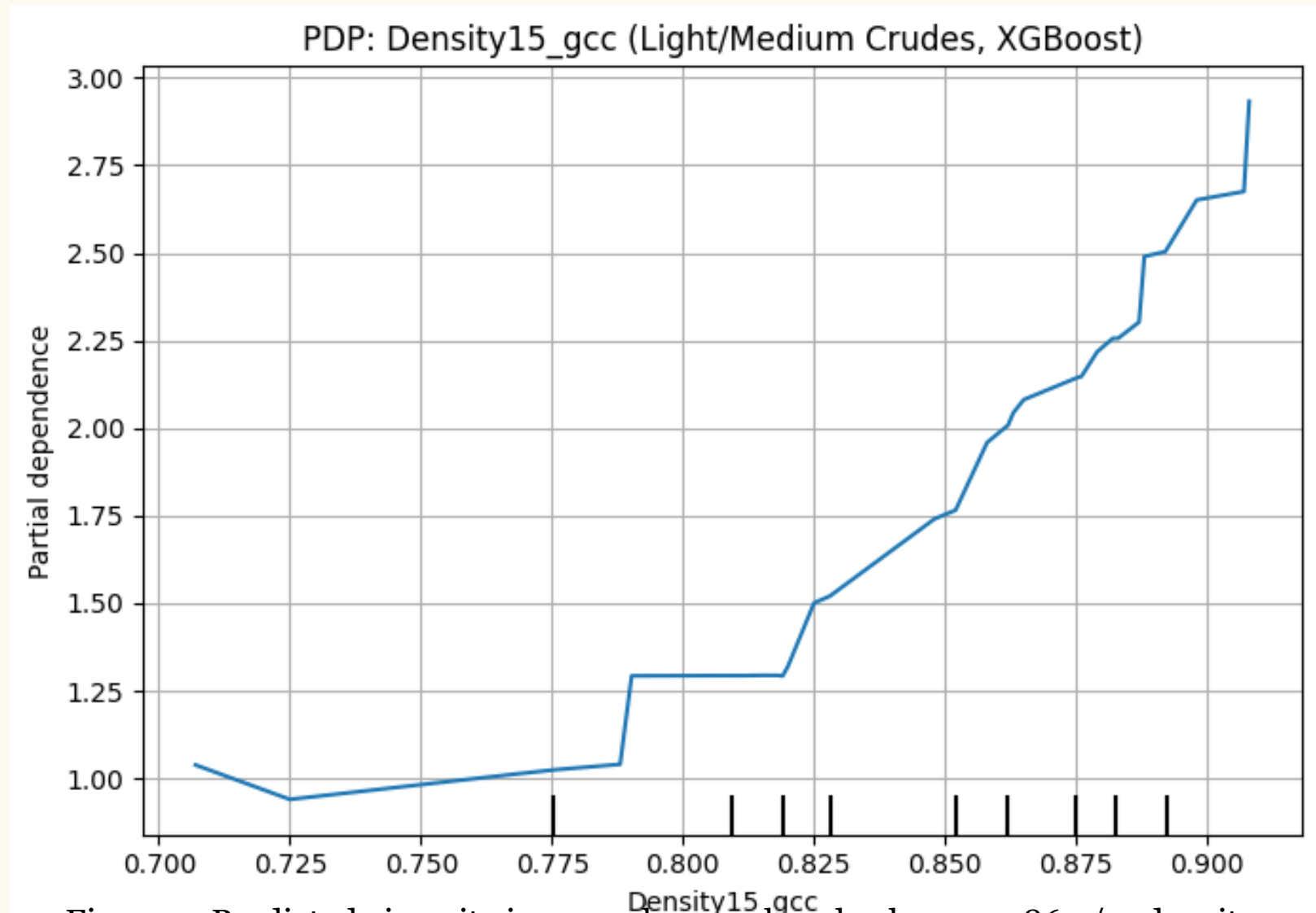


Figure 2. Predicted viscosity increased more sharply above ~0.86 g/cc density.

Discussion

Density consistently dominated across feature-importance, partial-dependence, and SHAP analyses. This was stronger than I initially expected.

The density trend is consistent with petroleum chemistry: denser crudes often contain larger, more complex molecules that resist flow. Aromatic content also showed a positive relationship with viscosity, while paraffin content showed a mild inverse relationship.

The model’s ability to capture these nonlinear patterns helps explain why XGBoost performed better than my initial linear-regression approach.

Conclusion

Using only five commonly available inputs, the model captured much of the viscosity variation in the light and medium crude oils studied.

This approach does not replace laboratory analysis, but it could provide faster preliminary estimates for operational decisions.

- Limitations:
- Limited number of samples - 52 samples
 - Light and medium crudes only
 - Viscosity evaluated only at 40 °C

Future work should test a larger dataset, include heavy crudes, and evaluate viscosity at additional temperatures.

References

- [1] ExxonMobil. Crude Oil Assays, 2024.
- [2] Equinor. "Crude Oil Assays."
- [3] Langeroudy Ahmady, K. P., et al. Scientific Reports, 2023.
- [4] Mohammed, R., et al. Processes, 2023.
- [5] Zrazhevsky, D., et al. Fuel, 2022.