Summary and quick context

This dataset contains 21,497 records of account/applicant phone and identity signals with a binary target "Fraud" (0 = no, 1 = yes). The data appears to combine identity scoring, underwriting scores, phone linkage checks, usage history and carrier/technology attributes — all signals commonly used to detect fraud at application or onboarding.

Below are the most important takeaways, each paired with the supporting table or chart from the notebook.

Headline findings

- Overall fraud prevalence is about 6.96% (1,497 fraudulent rows vs. 20,000 non-fraud).

  Supporting table:

Fraud	count
0	20000
1	1497

  And visualized class balance:

Download


Fraud class balance
20000
17500
15000
12500
10000
7500
5000
2500
0
count
0
1
Fraud

- The dataset is moderately imbalanced — not extreme, but enough to use precision/recall-aware metrics (PR-AUC, class weights, or threshold tuning) when modeling.

Data quality and missingness

- Most features are well populated, but two fields stand out for missingness:

column	missing_pct
Owner_VoIP_Indicator	82.51
Owner_MVNO_Indicator	18.25
Owner_Technology_Indicator	0.88
Owner_Parent_Phone_Carrier	0.88
Owner_Phone_Carrier	0.87
...	...

- Interpretation and action:
  - Owner_VoIP_Indicator is missing in ~82.5% of rows. That high missingness likely makes "is missing" itself a useful signal; add an is-missing flag rather than using the raw field directly.
  - Owner_MVNO_Indicator is missing in ~18%. Also worth treating missingness explicitly.
  - Nearly all other numeric features have very low missingness (<1%).

What numeric summaries reveal

- Numeric snapshot (selected columns):

	mean	std	min	25%	50%
IdentityScore	476.57	116.39	222.0	394.0	453.0
EAScore	319.20	247.55	21.0	92.0	235.0
UWScore	4.14	8.84	0.0	0.694	1.485
Owner_Verified_Components	3.40	2.31	1.0	2.0	2.0

  - Note: some variables (e.g., UWScore) have a heavy right tail or outliers (max far above 75th percentile). Several linkage fields contain negative codes (-1, -2) that are likely special codes (no match/unknown) rather than numeric negatives — treat these as categorical codes when modeling.

Which features relate to Fraud (directional associations)

- Correlation-ranked relationships (absolute magnitude, numeric features):

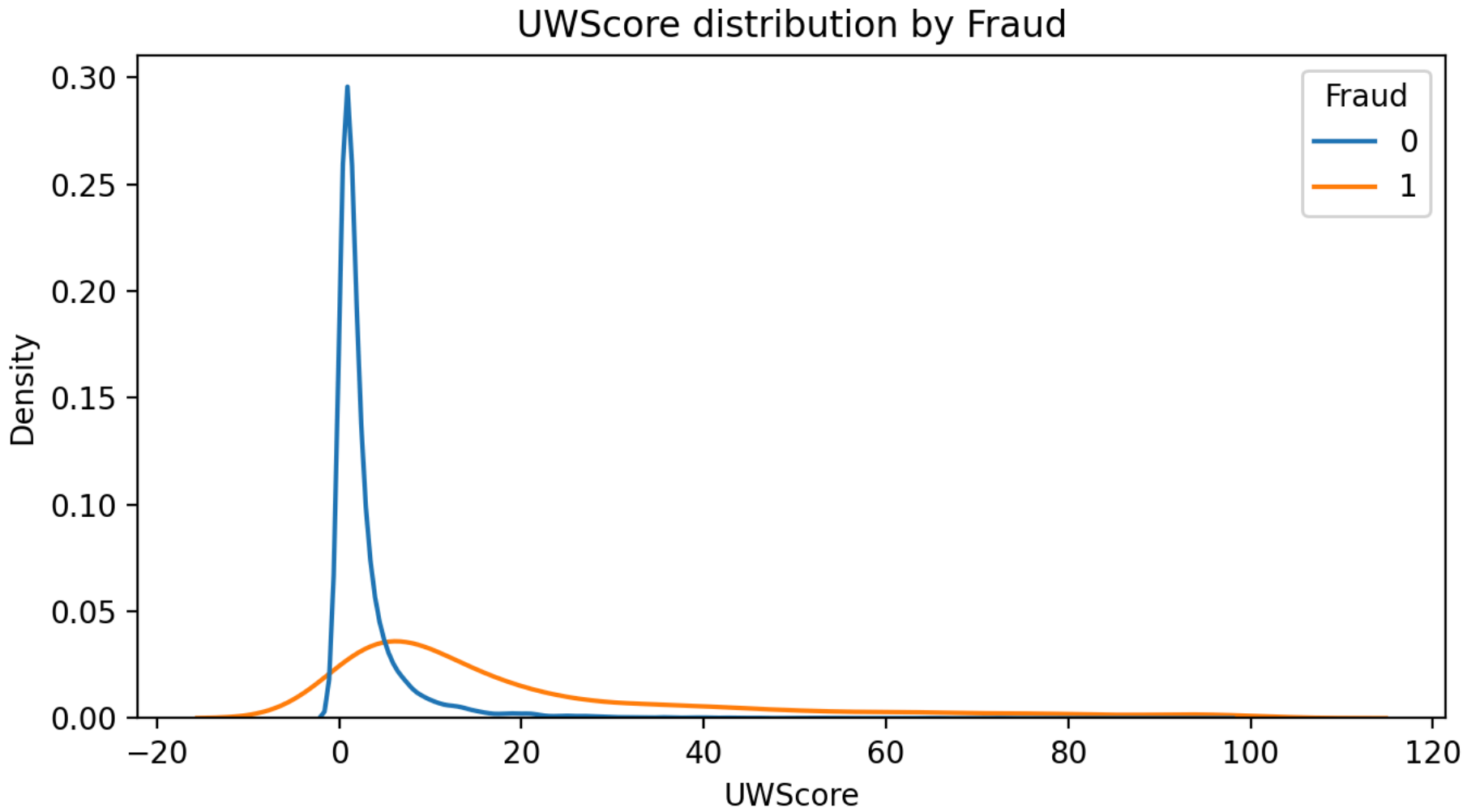
feature	correlation_with_Fraud
UWScore	0.49
IdentityScore	0.34
Identifier	0.34
Owner_Verified_Components	0.30

feature	correlation_with_Fraud
Owner_Phone_1_to_Name_Linkage	-0.30
Owner_Phone_Usage_Past_12_months	-0.29
Owner_Recent_Phone_Usage_Past_2_months	-0.26
Owner_VoIP_Indicator	-0.26
Owner_Address_to_Phone_1_Linkage	-0.17

    - Supporting correlation table: (excerpt shown above)
- Interpreting these patterns with reference plots:
    - UWScore is the single strongest numeric discriminator: fraud cases are shifted relative to non-fraud in the UWScore distribution.

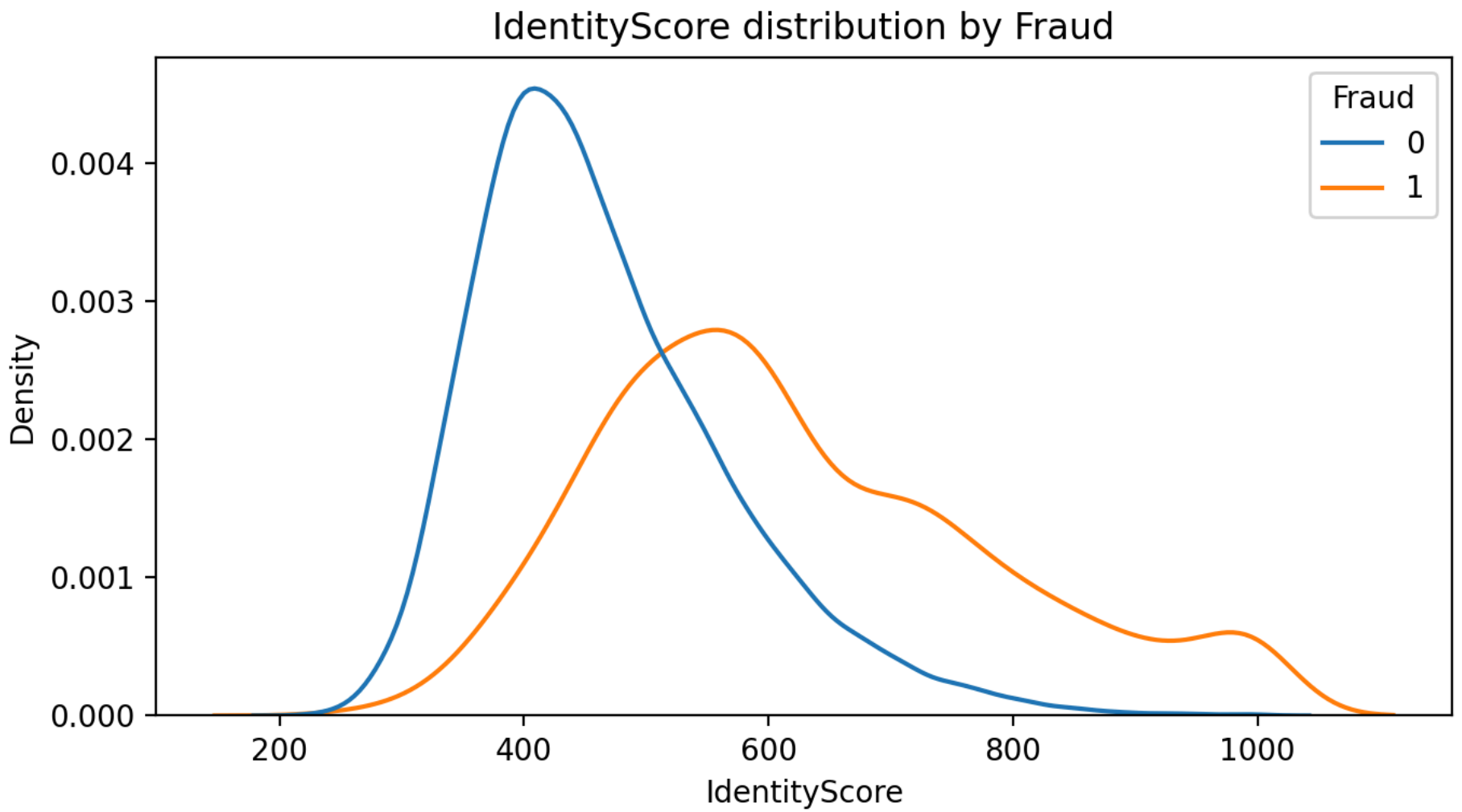
      See UWScore distribution by Fraud:

Download



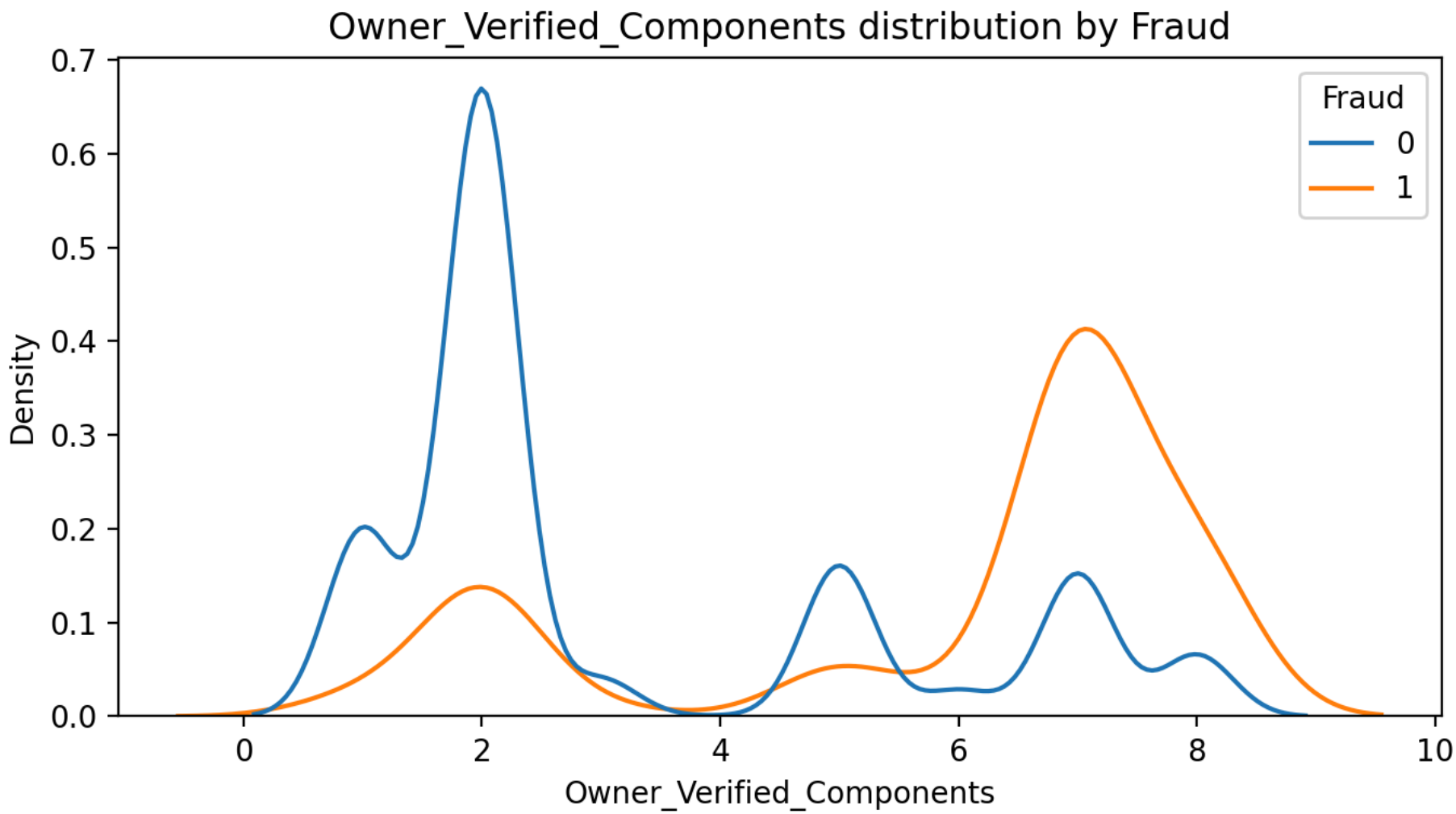
- IdentityScore also trends higher for fraud:

Download



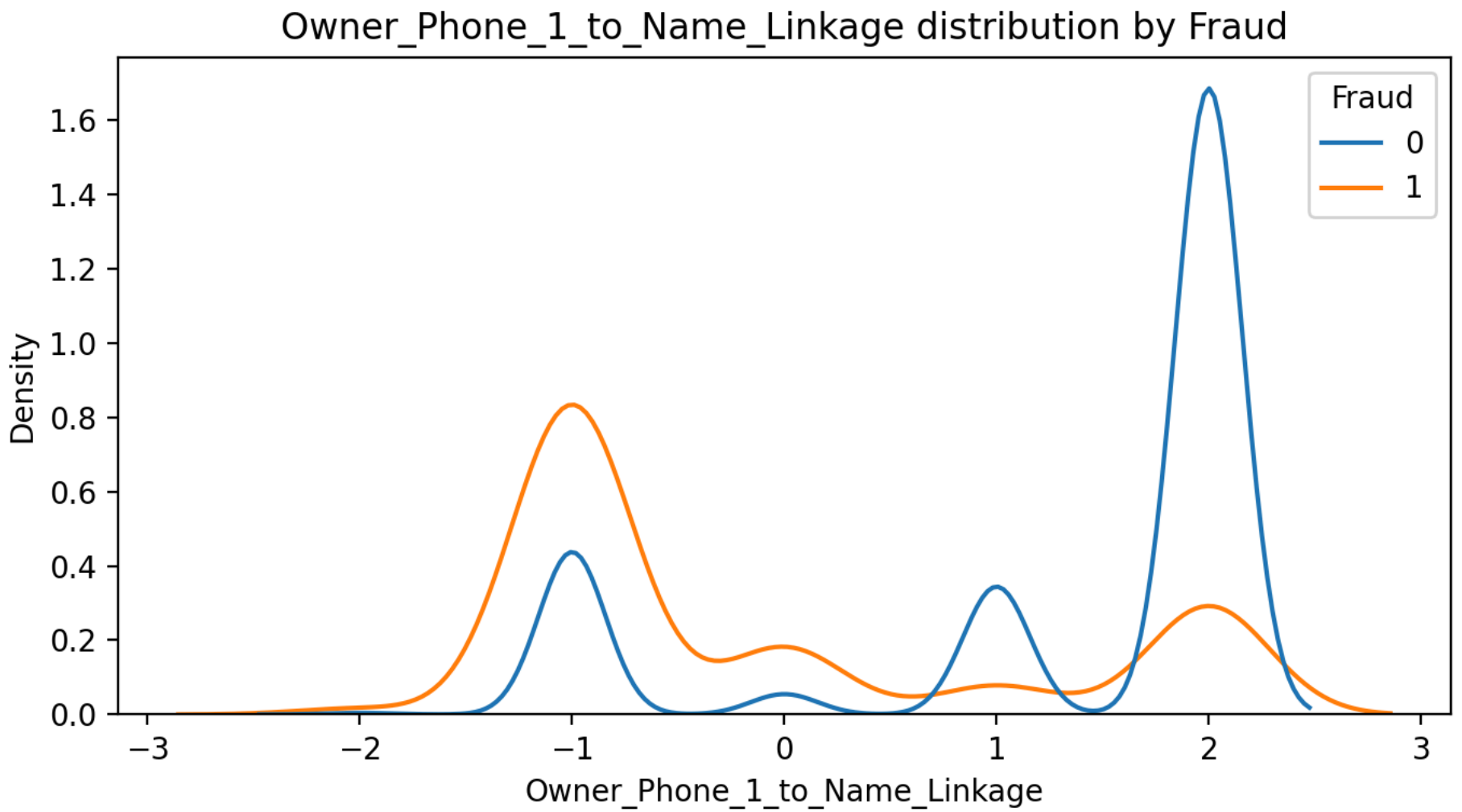
- Owner_Verified_Components is higher on average for fraud records (more verified components correlate with fraud here):

Download

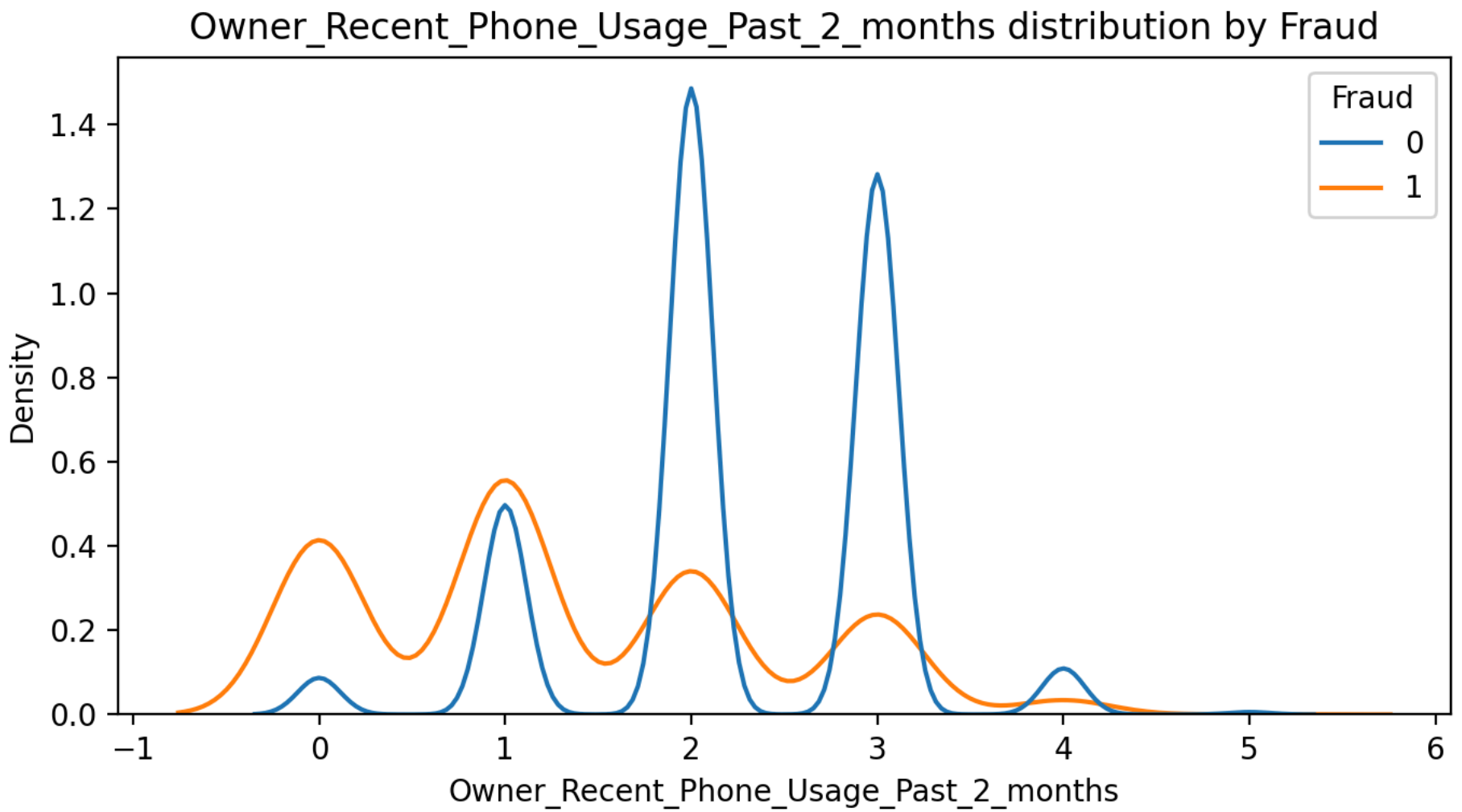


- Several phone linkage and usage metrics are lower (or have special-code mass) for fraud — e.g., phone-to-name linkage and recent usage. Visuals show fraud mass at lower/special-code values:

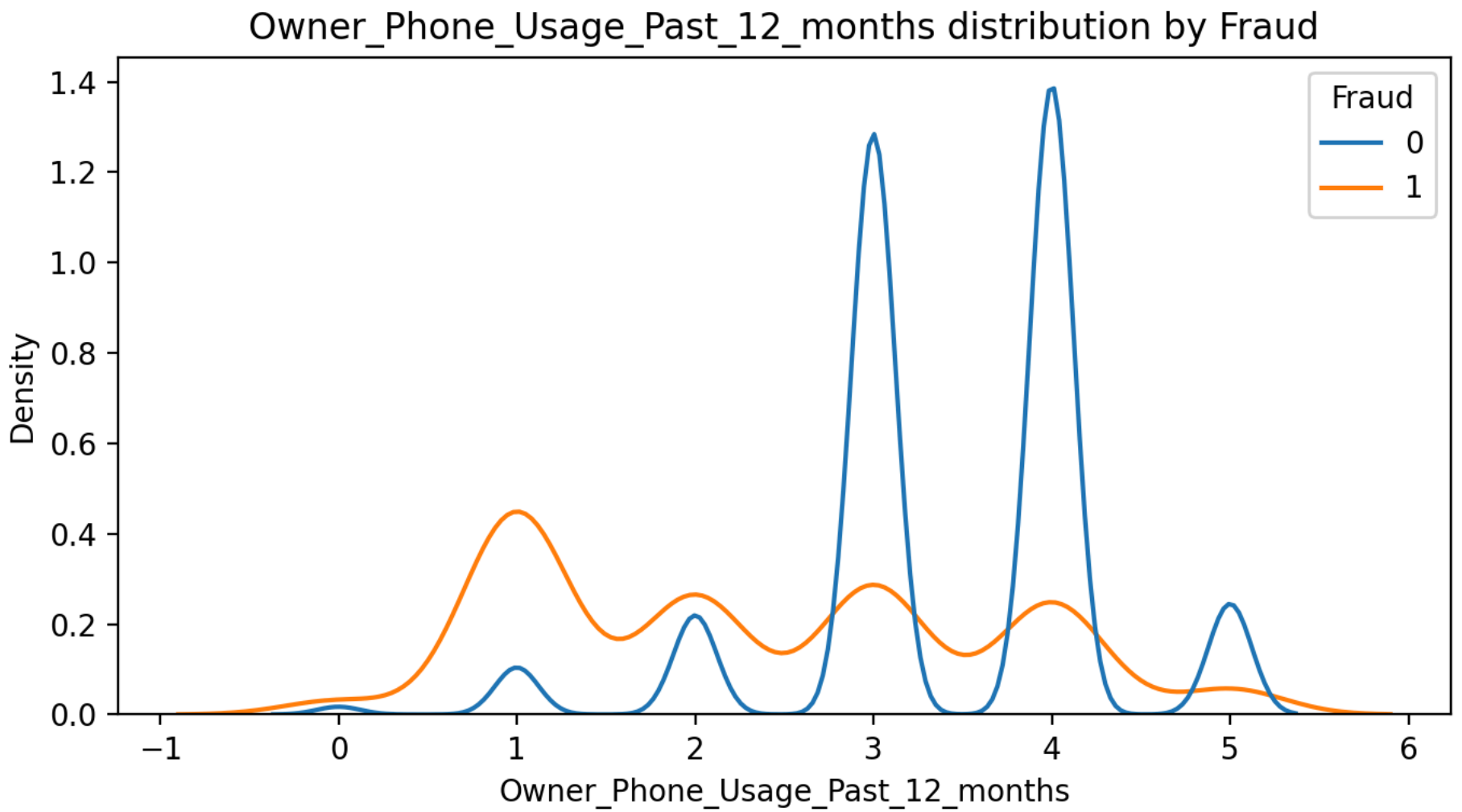
Download


Owner_Phone_1_to_Name_Linkage distribution by Fraud
Fraud
0
1
Density
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−3
−2
−1
0
1
2
3
Owner_Phone_1_to_Name_Linkage

Download


Owner_Recent_Phone_Usage_Past_2_months distribution by Fraud
Fraud
0
1
Density
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−1
0
1
2
3
4
5
6
Owner_Recent_Phone_Usage_Past_2_months

Download



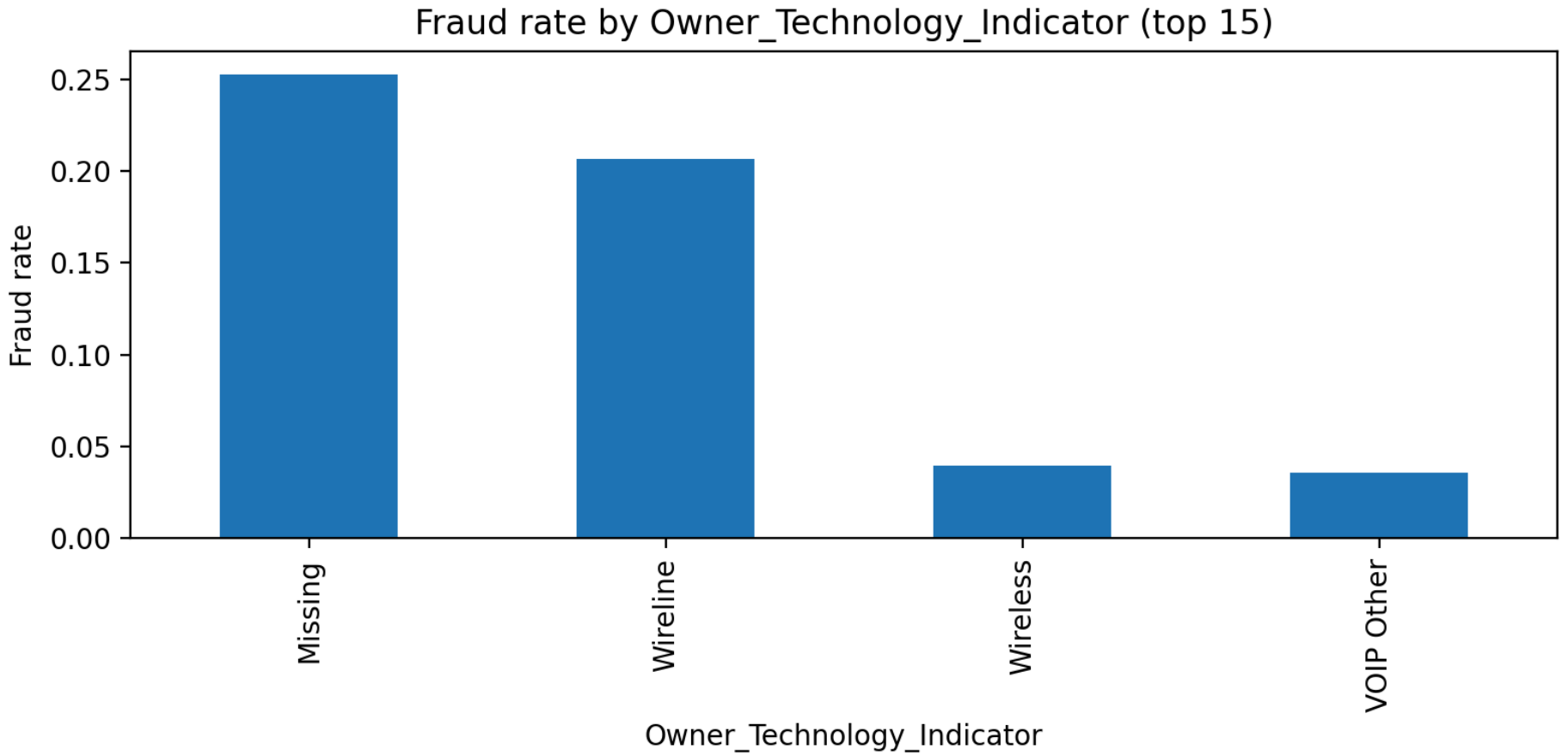
- Caution flag:
  - The Identifier field shows a notable correlation with Fraud. Identifiers should not be predictive; this suggests dataset ordering or process changes over time introduced leakage (e.g., batches with higher fraud). For reliable modeling, exclude Identifier or

investigate time/process confounding.

Categorical signals and carrier patterns

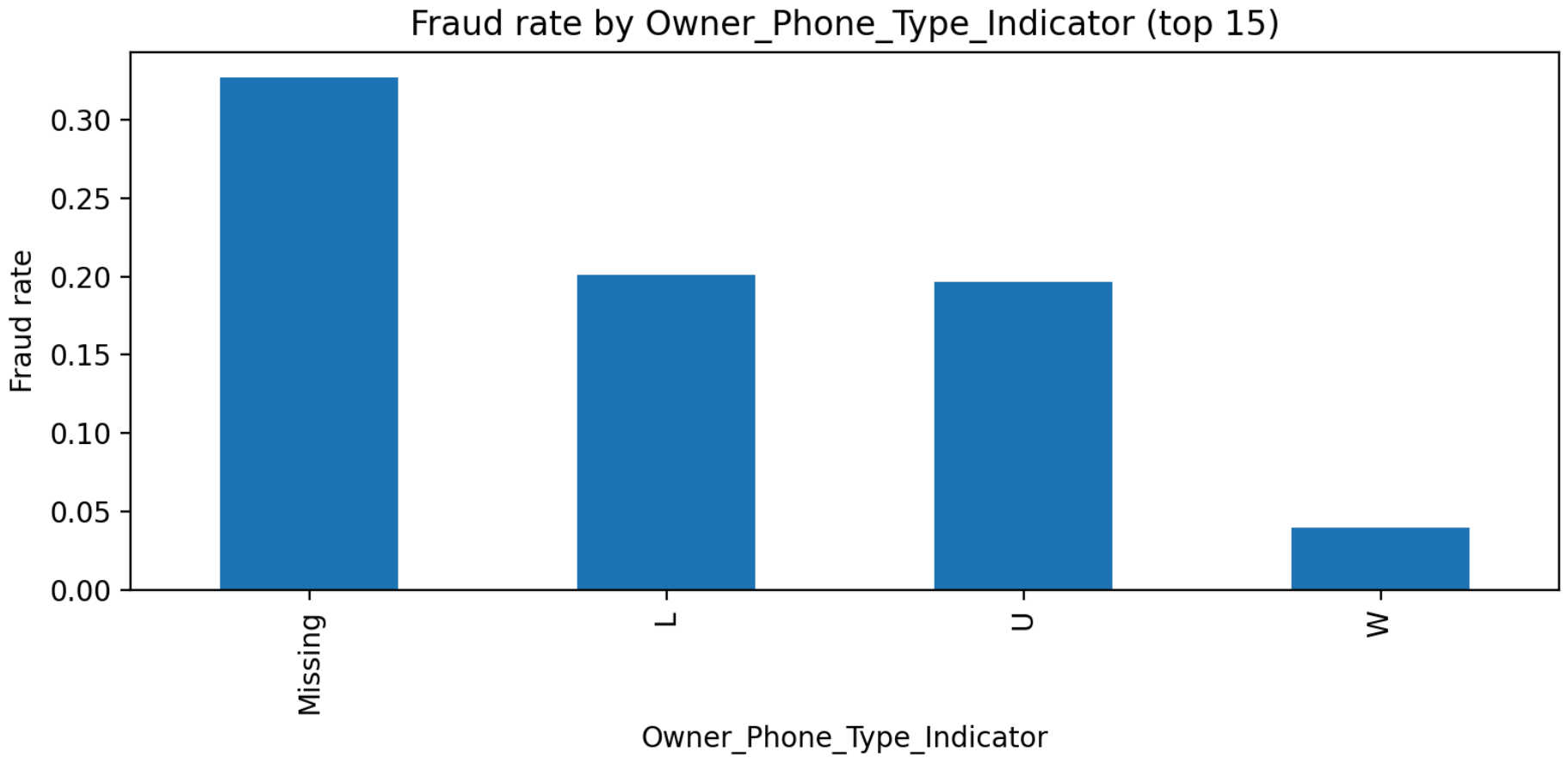
- Fraud rate by technology / phone type / prepaid / business / MVNO indicators are plotted; these show differences in fraud rate across categories (interpret with support). Example visuals:
  - Technology indicator fraud rates:

Download



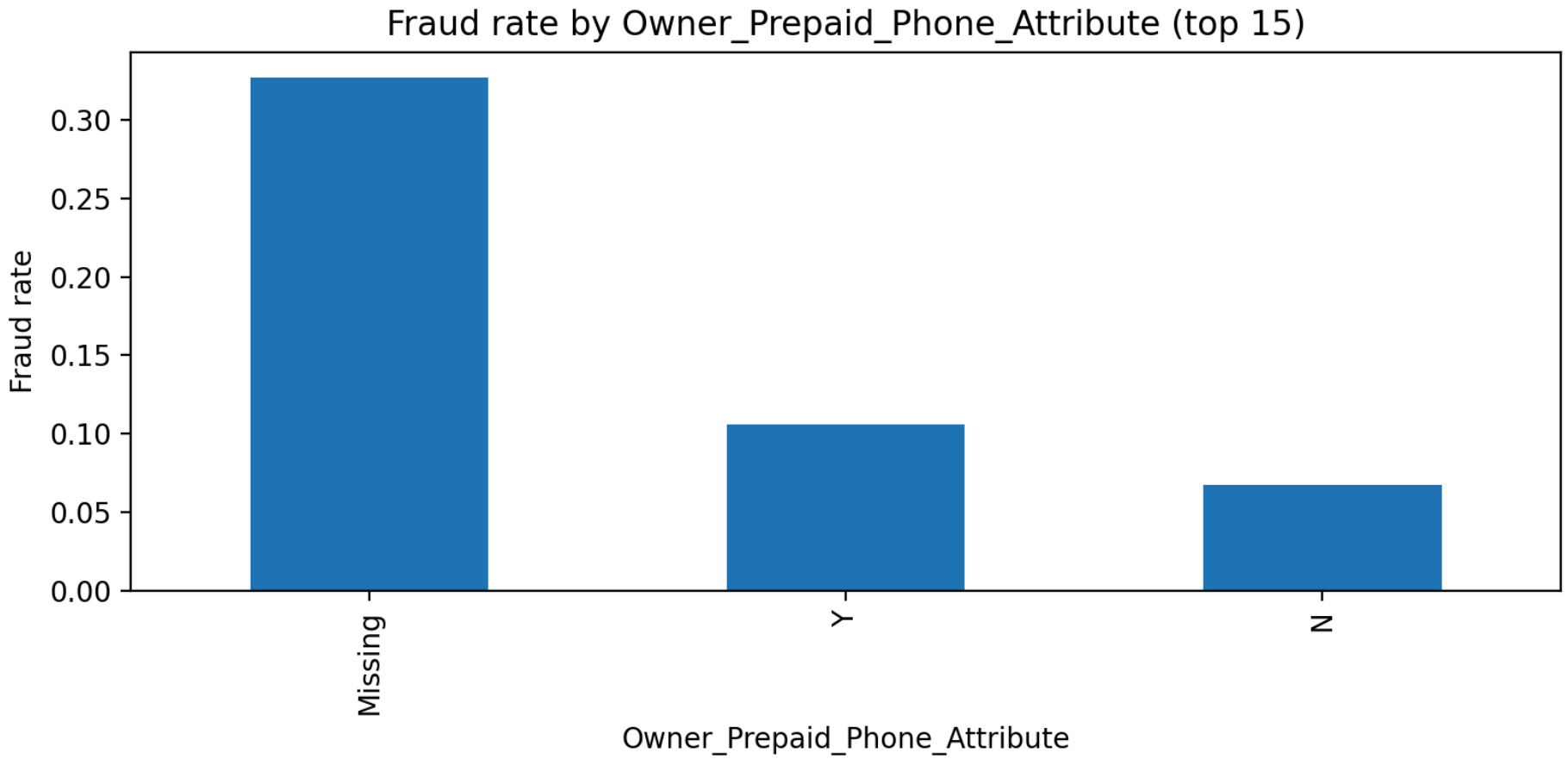
- Phone type indicator fraud rates:

Download



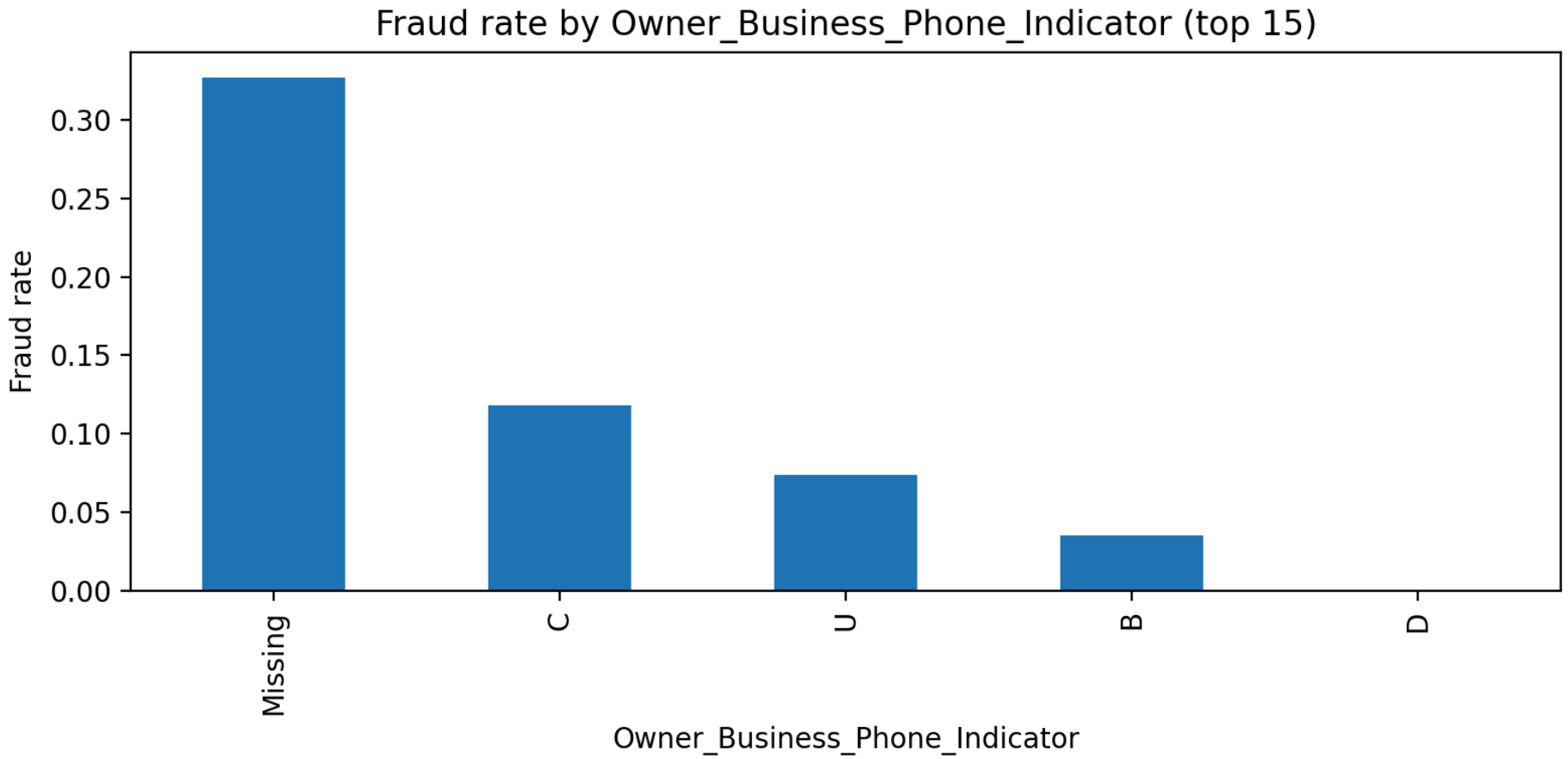
- Prepaid attribute fraud rates:

Download



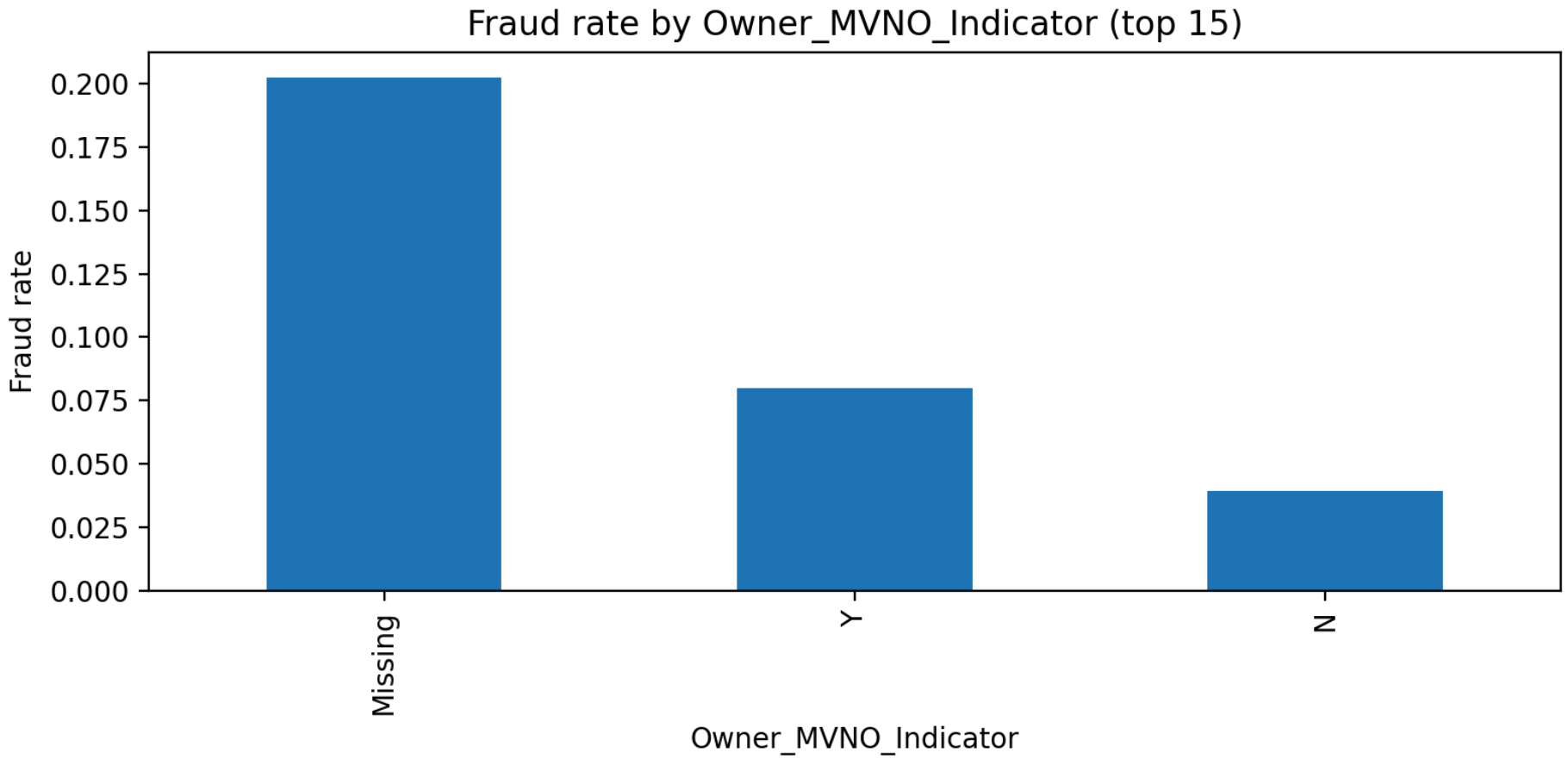
- Business phone indicator fraud rates:

Download



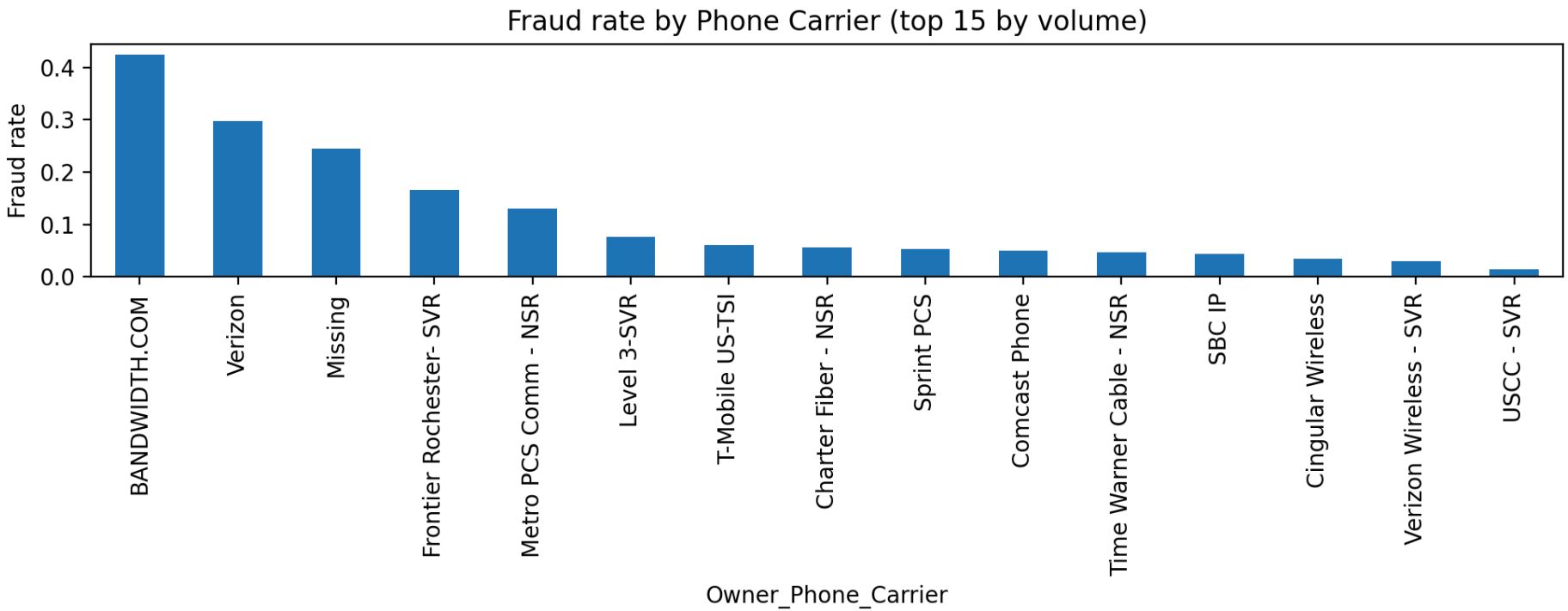
- MVNO indicator fraud rates:

Download



- Fraud rate by top 15 phone carriers:

Download



- Practical reading note: those charts show fraud *rates*, not counts. Categories with small counts can look extreme — pair rate with support when acting on these signals.

Practical recommendations

- For quick modeling or rule-setting:
  - Use UWScore, IdentityScore, Owner_Verified_Components, and phone linkage/usage features as primary signals.
  - Encode special codes (-1, -2) explicitly as categories; do not treat them as continuous numeric values.
  - Add missingness indicators for Owner_VoIP_Indicator and Owner_MVNO_Indicator (missing may itself be predictive).

    - Exclude Identifier to avoid leakage; investigate time/order effects that cause ID to correlate with fraud.
    - Evaluate models with PR-AUC and class-weighting or threshold tuning due to the ~7% fraud rate.
- For deeper business checks:
    - Produce fraud-rate-by-volume tables (rate + count + confidence intervals) for carriers and small categories before operationalizing rules.
    - Investigate why UWScore and IdentityScore are higher for fraud (score construction, thresholds, or upstream labeling artifacts).

Conclusion

- This dataset provides several clear signals that separate fraud from non-fraud: UWScore (strongest), IdentityScore, and several phone-linkage/usage variables. High missingness in Owner_VoIP_Indicator and partial missingness in Owner_MVNO_Indicator are themselves informative. Be careful about special-code values and the Identifier leakage. With careful encoding and evaluation focused on imbalanced metrics, you can build a reliable baseline fraud model using the highlighted variables.