1️⃣ Final Recommended Feature Set

Feature	Notes / Reason for Keeping
`FiberElectronicCombo`	Highest MI with target (0.15). Captures interaction between Internet type and payment method.
`PaymentMethod_Electronic`	High MI (0.145). Important standalone signal.
`Contract_TwoYear`	High MI (0.131). Predictive of churn.
`InternetService_Fiber`	Optional — moderate MI (0.12). Can drop to remove redundancy with combo feature.
`SeniorCitizen`	High MI (0.104). Independent demographic signal.
`TotalCharges`	Low MI alone (0.003), but contains financial info; kept for interpretability.

✅ **Lean, interpretable, and mostly independent. Minimal redundancy.**

2️⃣ Recommended Model / Ensemble

- **Soft-voting ensemble of:**
    1. **SVM (RBF)**
    2. **Gradient Boosting**
    3. **SVM (Linear)**
- **Pros: balances decision boundaries (SVMs) with tree-based model strengths (GBM).**
- **Use scaled features for SVMs, raw for GBM.**

3️⃣ Ensemble Performance (Fiber dropped)

Metric	Value

AUC	**0.915**
Accuracy	**0.855**
Precision	**0.832**
Recall	**0.890**
F1 Score	**0.860**

Interpretation:

- **These numbers are ideal for a lean model — you've removed redundancy, reduced overfitting, and kept strong predictive signals.**
- **AUC ~0.91–0.93 is very reasonable for churn datasets of this size.**

- **Recall 0.89 → excellent at catching churners (critical in business scenarios).**
- **Slight drops compared to the old 0.97 AUC are expected and actually indicate more realistic, generalizable performance.**

4️⃣ Practical Considerations

- **In production, these numbers are plausible and actionable.**
- **You could experiment with:**
  - **Small feature transformations (log scaling, binning)**
  - **Slight hyperparameter tuning of SVM/GBM**
  - **Weighted voting in the ensemble to favor recall**
- **But in practice, further “improvement” might only gain marginal 1–2% improvements, so the current setup is solid.**

Advanced Churn Prediction – Optimized Feature Set

This project demonstrates a churn prediction pipeline using a lean, interpretable feature set and an ensemble of machine learning models. The workflow emphasizes feature selection based on mutual information, redundancy reduction, and realistic model performance evaluation.

Feature Selection

The final feature set was chosen to maximize predictive power while minimizing redundancy:

Feature	Notes / Reason for Keeping
`FiberElectronicCombo`	Captures interaction between Internet type and payment method; highest MI with target.
`PaymentMethod_Electronic`	Strong standalone signal.
`Contract_TwoYear`	Predictive of long-term customer retention.

`InternetService_Fiber`	Optional; moderate MI. Can be dropped to reduce redundancy with `FiberElectronicCombo`.
`SeniorCitizen`	Demographic signal, independent of other features.
`TotalCharges`	Contains financial information; interpretable.

Ensemble Model

A soft-voting ensemble was constructed using:

1. SVM (RBF) – captures nonlinear decision boundaries
2. Gradient Boosting – tree-based, handles feature interactions
3. SVM (Linear) – adds linear perspective to ensemble

Note: Features were scaled for SVMs to ensure proper performance.

Performance (Optimized Feature Set, Fiber Dropped)

Metric	Value
AUC	0.915
Accuracy	0.855
Precision	0.832
Recall	0.890
F1 Score	0.860

Insights:

- **Performance is realistic and generalizable.**
- **Recall remains high, ensuring most churners are correctly identified.**

- **Ensemble balances the strengths of different models while avoiding overfitting caused by redundant features.**

Next Steps / Experimentation

- **Minor hyperparameter tuning for SVM/GBM may improve performance marginally.**
- **Optional feature transformations (log-scaling, binning) could be tested.**
- **Weighted voting could be applied to further prioritize recall if business objectives require it.**

If you like, I can also draft a small schematic diagram showing the ensemble and feature relationships for the README — it makes the repo much more visually appealing and easy to understand. Do you want me to do that?

Advanced Churn Prediction – Optimized Feature Set

This project demonstrates a **churn prediction pipeline** using a lean, interpretable feature set and an ensemble of machine learning models. The workflow emphasizes **feature selection based on mutual information**, redundancy reduction, and realistic model performance evaluation.

Feature Selection

The final feature set was chosen to maximize predictive power while minimizing redundancy:

Feature	Notes / Reason for Keeping
`FiberElectronicCombo`	Captures interaction between Internet type and payment method; highest MI with target.
`PaymentMethod_Electronic`	Strong standalone signal.
`Contract_TwoYear`	Predictive of long-term customer retention.
`InternetService_Fiber`	Optional; moderate MI. Can be dropped to reduce redundancy with `FiberElectronicCombo`.
`SeniorCitizen`	Demographic signal, independent of other features.
`TotalCharges`	Contains financial information; interpretable.

Ensemble Model

A **soft-voting ensemble** was constructed using:

1. **SVM (RBF)** – captures nonlinear decision boundaries
2. **Gradient Boosting** – tree-based, handles feature interactions
3. **SVM (Linear)** – adds linear perspective to ensemble

Note: Features were scaled for SVMs to ensure proper performance.

Performance (Optimized Feature Set, Fiber Dropped)

Metric	Value
AUC	0.915

Accuracy	0.855
Precision	0.832
Recall	0.890
F1 Score	0.860

Insights:

- Performance is realistic and generalizable.
- Recall remains high, ensuring most churners are correctly identified.
- Ensemble balances the strengths of different models while avoiding overfitting caused by redundant features.

Next Steps / Experimentation

- Minor hyperparameter tuning for SVM/GBM may improve performance marginally.
- Optional feature transformations (log-scaling, binning) could be tested.
- Weighted voting could be applied to further prioritize recall if business objectives require it.

If you like, I can also **draft a small schematic diagram showing the ensemble and feature relationships** for the README — it makes the repo much more visually appealing and easy to understand. Do you want me to do that?