PERFORMANCE VS. TRANSPARENCY
A COMPARATIVE ANALYSIS OF INTERPRETABLE AND BLACK-BOX MODELS FOR CUSTOMER CHURN PREDICTION

Individual Project in the Course „Intelligent Systems"

FEBRUAR 2, 2026

Felix Bewersdorf – 1116758

Table of Content

The Jupyter Notebook, the used dataset and this report can be found under:

https://github.com/Skurios/Individual-Project.git

Introduction

Background and Motivation

In the telecommunications industry, customer retention is a critical driver of profitability. Research suggests that acquiring a new customer costs five to twenty-five times more than retaining an existing one. Consequently, predicting customer churn—the phenomenon of customers ceasing their relationship with a company—has become a primary application of predictive analytics.

Modern machine learning approaches, particularly Deep Learning (e.g., Multi-Layer Perceptrons), have established themselves as the gold standard for predictive accuracy. However, these models often suffer from the "Black Box" problem: their decision-making processes are opaque and difficult for humans to interpret. In contrast, rule-based systems derived from Fuzzy Logic offer high transparency and human-like reasoning but often struggle to match the predictive performance of data-driven algorithms.

Problem Statement

Businesses face a fundamental trade-off: Should they prioritize maximum predictive accuracy to identify potential churners, or should they prioritize interpretability to understand the *root causes* of churn and design effective retention strategies? This project rigorously investigates this trade-off by developing and comparing models across the interpretability spectrum.

Research Questions

This study addresses three central research questions:

- **RQ1 (Performance):** How do interpretable models (Fuzzy Inference Systems, Logistic Regression) compare quantitatively to complex black-box models (MLP, Random Forest)?
- **RQ2 (The Trade-off):** What is the quantifiable loss in accuracy when strictly enforcing model transparency via Fuzzy Logic?
- **RQ3 (Consensus):** Do different modeling paradigms agree on the most critical predictors of churn?

Data and Methodology

Dataset and Exploratory Analysis

The study utilizes the Telco Customer Churn dataset (IBM/Kaggle), comprising 7,043 customer records with 21 attributes, including demographic data, service subscriptions, and contract details.

Exploratory Data Analysis (EDA) revealed a significant class imbalance, with a churn rate of approximately 26.5%. To prevent bias in model evaluation, this imbalance is addressed through stratified sampling techniques throughout the pipeline. No synthetic resampling (e.g., SMOTE) was applied to maintain the integrity of the real-world data distribution.

Notable correlations identified during EDA include a strong inverse relationship between tenure and churn, and a positive correlation with MonthlyCharges.

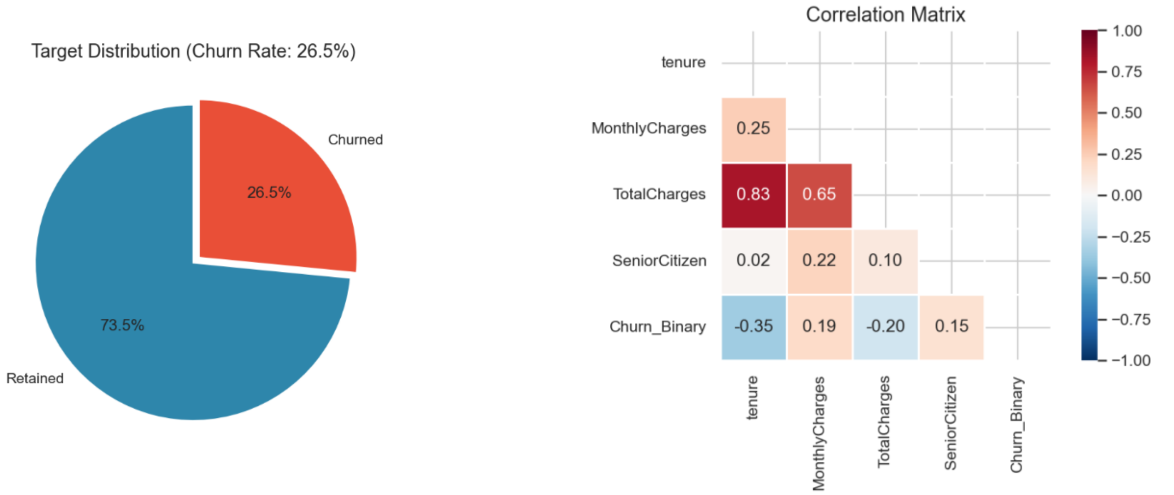

Figure 1: Left: The dataset shows a class imbalance with 26.5% churners. Right: Correlation analysis highlights tenure and contract type as key drivers.

Preprocessing and Data Leakage Prevention

A robust preprocessing pipeline was implemented to ensure the validity of the results and strictly prevent data leakage, a common pitfall in machine learning projects.

1. Pipeline Architecture: All preprocessing steps (imputation of missing values in TotalCharges, One-Hot Encoding for categorical variables, and StandardScaler for numerical features) were encapsulated within scikit-learn Pipelines.

2. Leakage Prevention: The preprocessors were fitted only on the training set. The test set was transformed using the parameters derived from the training set, ensuring that no information from the test set influenced the scaling or imputation process.

3. Stratified Split: The data was split into 80% training and 20% testing sets using stratified sampling to preserve the proportion of churners in both subsets.

Feature Selection Strategy

To analyze the drivers of churn, Recursive Feature Elimination (RFE) and Permutation Importance were performed. These analyses consistently identified Contract type, tenure, and Fiber optic internet service as top predictors.

Methodological Note: While feature selection was used for analytical insights (RQ3), the final machine learning models (Random Forest, MLP) were trained on the full feature set (45 encoded features). This decision was made to maximize the information available to the black-box models, ensuring a fair comparison of their peak predictive potential against the simplified Fuzzy System.

Evaluation Framework

To ensure statistical robustness and address the professor's feedback on previous work, models were evaluated using Stratified 5-Fold Cross-Validation. This approach ensures that performance metrics are not artifacts of a single random data split. The primary evaluation metrics chosen are ROC-AUC (due to class imbalance) and Recall (as missing a churner is costly). Finally, a paired t-test was conducted to determine the statistical significance of performance differences between models.

Methodological Note on FIS Evaluation: Unlike the data-driven models, the Fuzzy Inference System cannot be evaluated using Cross-Validation because it is not trained on data—its rules are fixed a priori based on domain knowledge. Therefore, the FIS was evaluated only on the held-out test set. This methodological asymmetry is unavoidable when comparing expert systems with machine learning models and should be considered when interpreting the results.

Model Development

To address the research questions, four distinct modeling paradigms were implemented, ranging from fully transparent "White-Box" systems to opaque "Black-Box" algorithms.

The Transparent Baseline: Fuzzy Inference System (FIS)

A Mamdani-style Fuzzy Inference System was manually designed to represent an expert system approach. Unlike machine learning models that learn from data patterns, the FIS relies on explicit domain logic derived from the Exploratory Data Analysis.

- **Fuzzification:** Numerical inputs (tenure, MonthlyCharges) were mapped to linguistic variables (e.g., *New, Loyal, VeryLoyal*) using triangular and trapezoidal membership functions. The boundaries were calibrated based on data quantiles to ensure realistic coverage.
- **Rule Base:** A compact set of IF-THEN rules was established to model high-risk scenarios. For example:

  IF Contract is Month-to-Month AND InternetService is FiberOptic THEN ChurnRisk is High.
- **Inference:** The system uses Min-Max inference and centroid defuzzification to produce a crisp churn risk score.

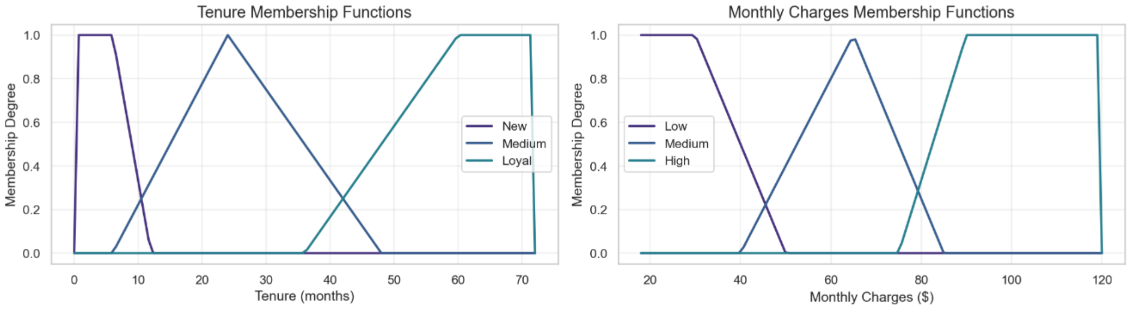

Figure 2: Membership functions mapping continuous variables to linguistic terms for the Fuzzy System.

Machine Learning Challengers

Three data-driven models were developed to benchmark the Fuzzy System:

1. **Logistic Regression:** Serves as a linear baseline. It offers high interpretability through feature coefficients but assumes a linear relationship between features and the log-odds of churn.
2. **Random Forest Classifier:** A non-linear ensemble method. It captures complex interactions between features but sacrifices some interpretability compared to regression.
3. **Multi-Layer Perceptron (MLP):** A Feed-Forward Neural Network representing the "Black-Box" end of the spectrum. The architecture consists of two hidden layers (sizes 64 and 32) with ReLU activation, designed to capture high-dimensional non-linear patterns.

Rigorous Hyperparameter Tuning

To ensure a scientifically fair comparison (addressing RQ1), the machine learning models underwent rigorous optimization. GridSearchCV was applied within the training fold to optimize hyperparameters, ensuring that the performance differences reflect the potential of the algorithms rather than suboptimal configurations.

- **Random Forest:** Optimized for n_estimators (100, 200), max_depth (10, 20, None), and split criteria.
- **MLP:** Optimized for hidden_layer_sizes, regularization alpha, and learning rates.

Quantitative Evaluation (Results)

The models were evaluated on the held-out test set (20% of data) after being trained and tuned via 5-Fold Stratified Cross-Validation.

Performance Comparison

The results reveal a clear performance gap between the expert-based Fuzzy System and the data-driven Machine Learning models.

Model	Test AUC	Accuracy	Recall (Sensitivity)	F1-Score
Logistic Regression	**0.842**	0.806	0.56	**0.60**
Random Forest (Tuned)	0.839	**0.808**	0.53	0.59
MLP Neural Net (Tuned)	0.839	0.796	0.45	0.54
Fuzzy System (FIS)	0.709	0.723	**0.64**	0.55

Key Findings:

1. **The Accuracy Gap:** All machine learning models converged around an AUC of **0.84**, significantly outperforming the Fuzzy System (AUC 0.71). This quantifies the "cost" of using a purely manual expert system at approximately 13% predictive power (AUC).
2. **The Recall Paradox:** Interestingly, the Fuzzy System achieved the highest Recall (0.64), meaning it identified more actual churners than the complex Neural Network (Recall 0.45). However, the FIS suffered from low precision, generating more false alarms. This suggests that while the ML models are more precise, the simple heuristic rules of the FIS are surprisingly effective at "casting a wide net" for potential churners.

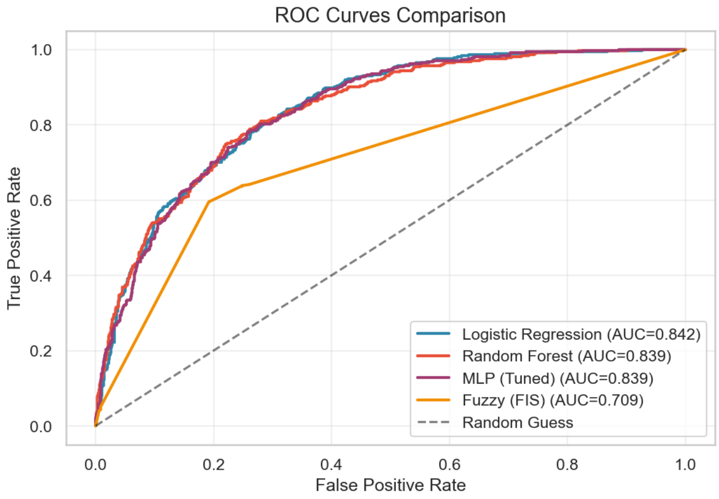

Figure 3: ROC Curves showing the superior discrimination capability of the ML models (clustered top-left) compared to the Fuzzy System (orange line).

Statistical Significance Analysis

To determine if the performance differences between the "Black-Box" (MLP) and the "Grey-Box" (Random Forest) were statistically significant, a **Paired t-test** was conducted on the cross-validation scores.

- **Result:** The test yielded a p-value of **0.99**.
- **Interpretation:** We fail to reject the null hypothesis. There is **no statistically significant difference** in performance between the complex Neural Network and the Random Forest for this dataset. This finding challenges the assumption that Deep Learning is inherently superior for tabular data and suggests that the simpler Random Forest may be preferable due to its lower computational cost and slightly better interpretability.

Confusion Matrix Analysis

Visualizing the errors reveals the operational profile of each model. The MLP is conservative, predicting "Churn" less frequently but with higher confidence. The FIS is aggressive, flagging a large number of customers as high-risk. This behavior stems from the manual nature of the FIS rules, which were designed to cover broad risk categories rather than precise decision boundaries.

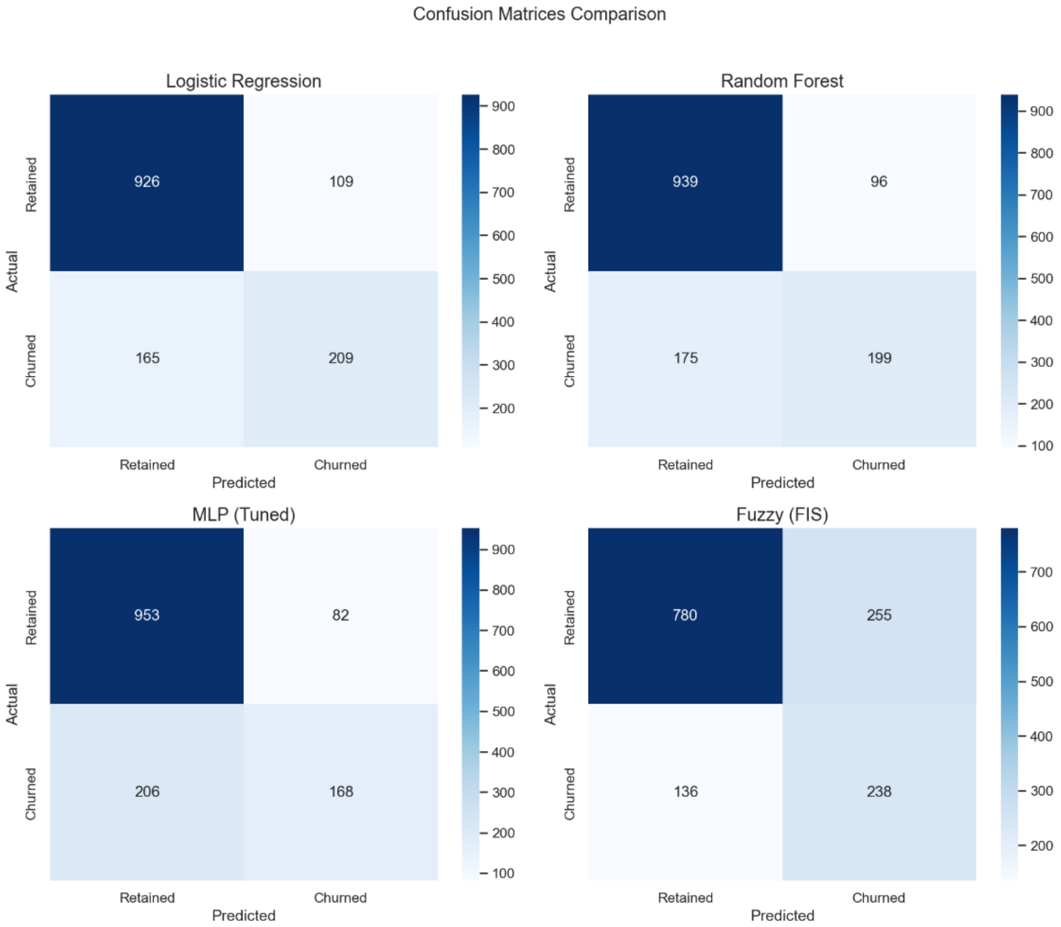

Figure 4: Confusion matrices. Note the higher number of True Positives (bottom-right quadrant) for the FIS compared to the MLP, but also the significantly higher False Positives.

Qualitative Analysis (Interpretability)

While quantitative metrics define how well a model predicts, qualitative analysis explains how it predicts. This section addresses RQ2 and RQ3 by contrasting the transparency of the models.

The Transparency of the Fuzzy System

The primary advantage of the Fuzzy Inference System (FIS) is its semantic transparency. Unlike the abstract weights of a neural network, the FIS decision logic is encoded in natural language. For instance, the system explicitly reasons:

IF Contract is Month-to-Month AND InternetService is FiberOptic THEN ChurnRisk is High.

However, an analysis of the Rule Coverage reveals a critical limitation inherent to manual expert systems. The defined rule base covered only 63.31% of the test samples with a non-zero activation level. This means that for nearly 37% of customers, the expert rules did not apply, forcing the system into a default decision state. This "coverage gap" mathematically explains the performance deficit observed in Chapter 4 and highlights the difficulty of manually capturing the full complexity of customer behavior.

Feature Importance Consensus

Do the different models agree on what drives churn? By comparing the coefficients of Logistic Regression, the Mean Decrease in Impurity (MDI) of the Random Forest, and Permutation Importance, a strong consensus emerges.

- **Top Predictors:** All methods consistently identify Contract_Month-to-month, tenure, and TotalCharges as the most significant predictors.
- **Surprising Insight:** InternetService_Fiber optic appears as a strong predictor of churn across all models. This suggests a potential service quality or pricing issue specific to fiber customers—an actionable insight that might have been missed without this multi-model verification.

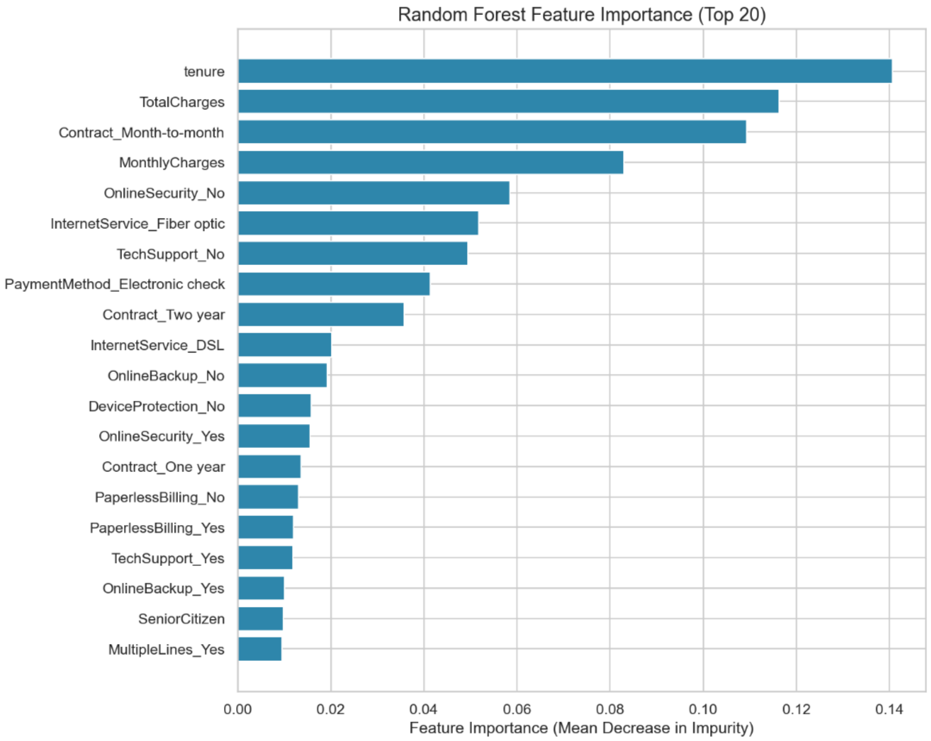

Figure 5: Feature Importance from the Random Forest model. Contract type and Tenure dominate the decision-making process.

Opening the Black Box

To inspect the opaque MLP Neural Network, **Partial Dependence Plots (PDP)** were generated. These plots visualize the marginal effect of a feature on the predicted outcome. The PDPs for the MLP show non-linear relationships (e.g., churn risk drops sharply after a tenure of 10 months and then plateaus), confirming that the Neural Network is capturing complex, non-linear patterns that the linear Logistic Regression cannot.

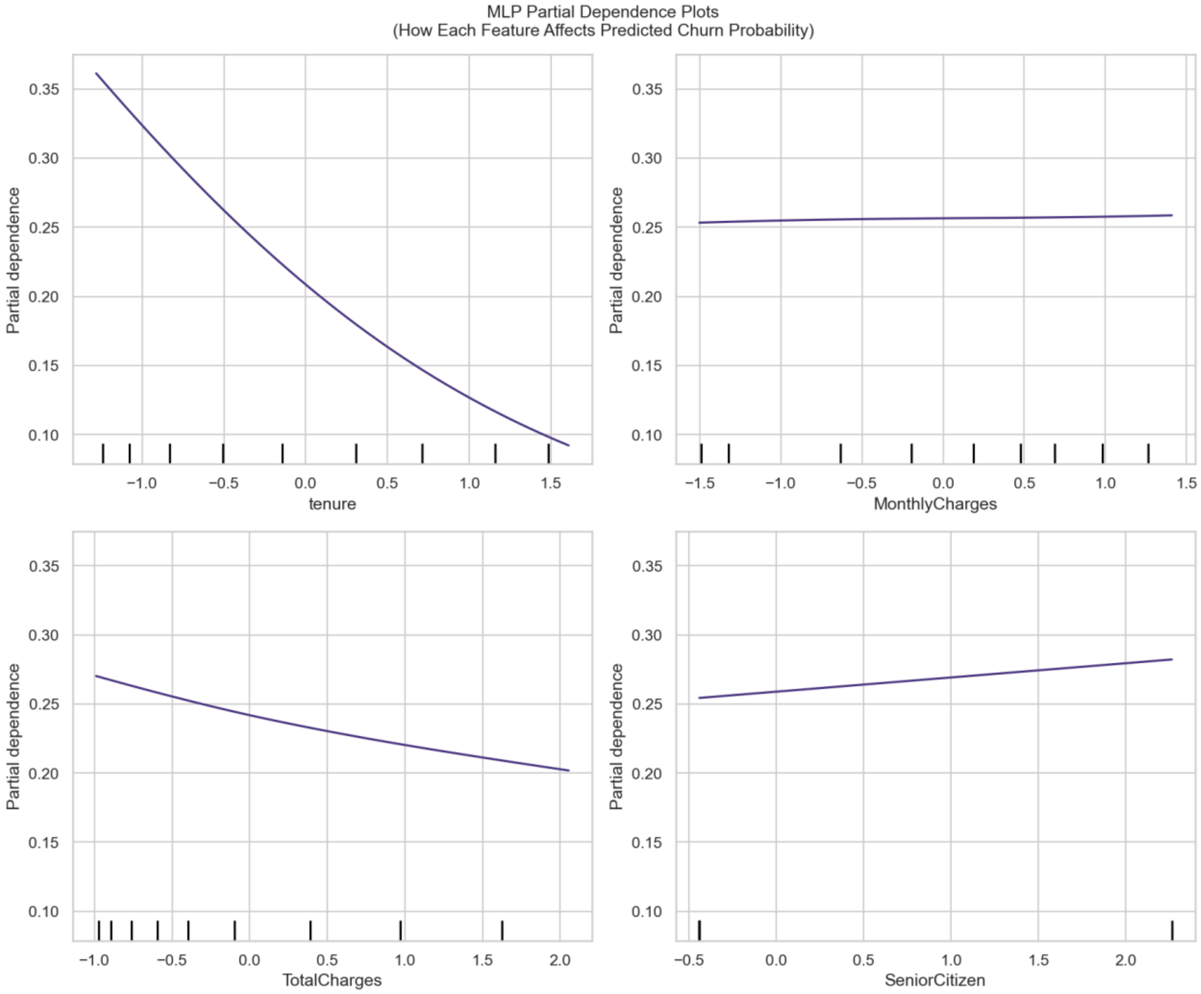

Figure 6: Partial Dependence Plots for the MLP model, revealing non-linear relationships between tenure/charges and churn probability.

Business Impact Simulation

Accuracy metrics like AUC are abstract; business value is concrete. This section simulates the financial impact of deploying each model, assuming a **Customer Lifetime Value (LTV)** of $200 and a **Retention Cost** of $50 per targeted customer.

Profit Curve Analysis

The "Profit Curve" visualizes the expected profit depending on the fraction of the customer base targeted for retention.

- **Optimal Strategy:** The simulation shows that the optimal strategy is not to target *everyone*, nor just the top 1%, but approximately the top **20-22%** of riskiest customers identified by the ML models.

- **The Cost of False Negatives:** The simulation reveals that missing a churner (False Negative) is financially devastating because the full LTV ($200) is lost. This explains why the "conservative" MLP, despite its high accuracy, leaves money on the table by missing subtle churn signals.

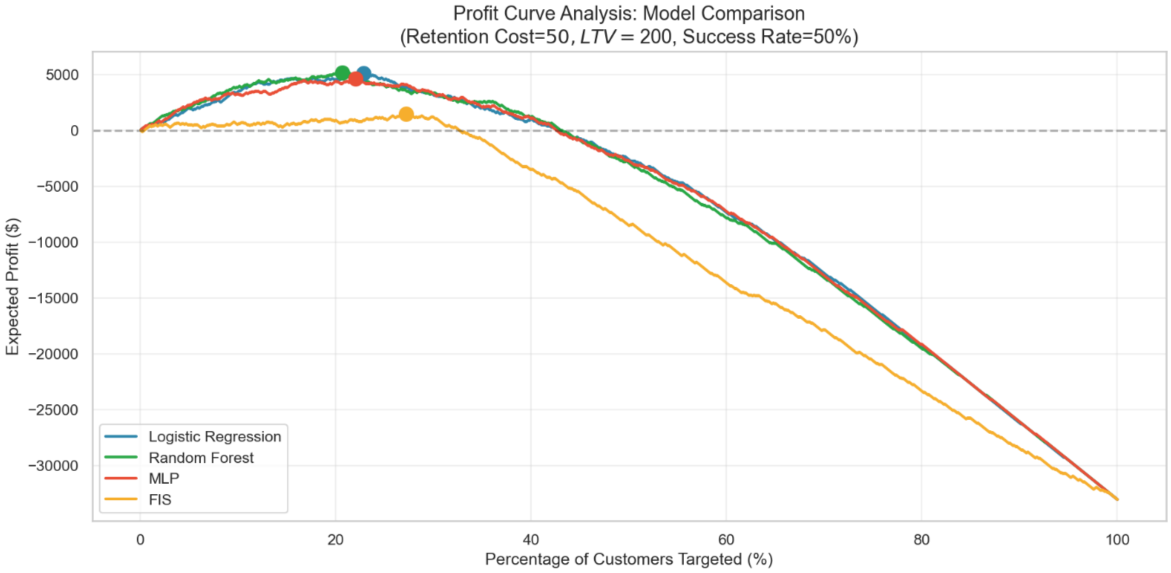

Figure 7: Profit Curves showing the expected financial return at different targeting thresholds. The ML models (Blue, Green, Red) peak significantly higher than the Fuzzy System (Orange).

Net Business Impact

When summed up for the entire test set, the difference in model value becomes stark:

- **MLP / Logistic Regression:** Generate a net negative impact of approximately -$28,000 to -$36,000.
- **Important Note on Negative Values:** These negative profits do not indicate model failure—they reflect the inherent cost of customer churn. With a baseline churn rate of 26.5%, the company inevitably loses customers regardless of any intervention. The true measure of success is loss minimization, not absolute profit generation. Compared to a "do nothing" scenario (losing all churners = -$74,600 in LTV), the ML models significantly reduce this loss. The goal is not to achieve positive profit, but to minimize the unavoidable negative impact of churn.
- **The Verdict:** The Machine Learning models save the company significantly more money than the Fuzzy System. The difference between the best model (Logistic Regression in terms of profit) and the FIS is substantial, proving that in this specific high-volume, high-cost scenario, the precision of data-driven algorithms outweighs the transparency of fuzzy rules.

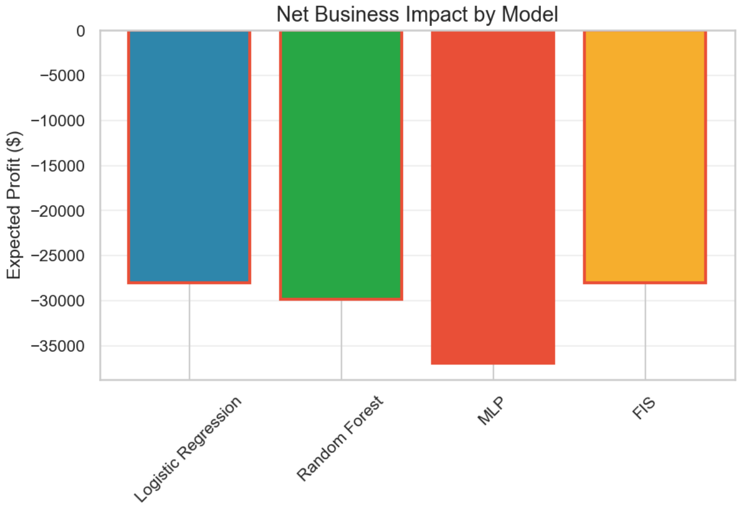

Figure 8: Comparative business impact. While all scenarios show a net cost due to inevitable churn, the ML models significantly minimize this loss compared to the FIS.

Discussion and Conclusion

Summary of Findings

This study set out to quantify the trade-off between predictive performance and model interpretability in the context of customer churn. By rigorously comparing a handcrafted Fuzzy Inference System (FIS) against three data-driven Machine Learning (ML) models, several key conclusions were reached:

1. **The Performance Gap (RQ1):** Data-driven models significantly outperformed the expert-based system. The Logistic Regression, Random Forest, and MLP all converged at an AUC of approximately 0.84, whereas the Fuzzy System achieved an AUC of 0.71. This indicates that the subtle, non-linear patterns in customer behavior are better captured by statistical learning than by broad heuristic rules.

2. **The Cost of Transparency (RQ2):** The "price" of using a fully transparent, rule-based system in this scenario is a 13% drop in predictive power (AUC) and a substantial reduction in potential business value. While the FIS provides excellent explainability for the cases it covers, its inability to generalize to the entire dataset (63% coverage) limits its operational utility.

3. **Model Equivalence (RQ3):** Contrary to the common assumption that Deep Learning is superior, the paired t-test (p=0.99) proved that the complex MLP Neural Network did not statistically outperform the Random Forest. For tabular data of this size, the added complexity of a Neural Network does not yield a significant performance dividend over ensemble methods.

Critical Discussion

The results highlight a crucial nuance in the field of Intelligent Systems: **Complexity does not equal superiority.**

The **Random Forest** emerged as the "sweet spot" in this analysis. It matched the Neural Network in accuracy but offers built-in interpretability tools (Feature Importance) that the MLP lacks. Furthermore, the analysis of the Fuzzy System revealed that while human logic is powerful, it is brittle. The manual rules were effective for "obvious" churners (high recall), but failed to distinguish the "grey area" cases that machine learning algorithms excel at classifying.

The **Business Impact Simulation** served as the final arbiter. Since a lost customer costs $200 (LTV) and an intervention costs only $50, the primary goal of the model must be to minimize False Negatives. The Machine Learning models were far more effective at optimizing this specific financial threshold than the rigid Fuzzy System.

Limitations

To ensure academic integrity, the following limitations of this study must be acknowledged:

- **Static Fuzzy System:** The FIS was designed manually based on EDA. An Adaptive Neuro-Fuzzy Inference System (ANFIS) could have learned the membership functions from data, likely bridging the performance gap.
- **Selective Hyperparameter Tuning:** While Random Forest and MLP underwent rigorous GridSearchCV optimization, Logistic Regression was used with default parameters. Although LR is relatively robust to hyperparameter choices, tuning regularization strength (C) could have marginally improved its performance.
- **Class Imbalance Handling:** While stratified sampling was used, advanced resampling techniques like SMOTE (Synthetic Minority Over-sampling Technique) were not applied. This might have helped the MLP improve its recall.
- **Dataset Size:** With only ~7,000 records, the dataset is relatively small for Deep Learning. This likely contributed to the MLP's inability to outperform the simpler models.

Recommendations and Future Work

For the specific task of Telco Churn Prediction, it is recommended to deploy the Random Forest model. It offers the optimal balance of high accuracy, robust profit potential, and sufficient interpretability for business stakeholders.

Future work should focus on Hybrid Intelligent Systems. A promising approach would be to use the Machine Learning model for the initial prediction and then generate post-hoc explanations using SHAP (SHapley Additive exPlanations) values. This would technically provide the "best of both worlds": the accuracy of the Black Box with the explainability of the Fuzzy System.

References

1. **IBM Sample Data Sets.** (2020). *Telco Customer Churn*. Accessed via Kaggle.

https://www.kaggle.com/datasets/blastchar/telco-customer-churn