DS 325 Final Project Report

DS 325B Spring 2025

Professor Eatai Roth

Salmin Mwinjuma, Thanh (Ethan) Nguyen

DS 325 Final Project Report

I/ Introduction/Abstract

Customer churn is a major revenue risk for telecom providers, especially in the business segment where contracts are high-value and long-term. Traditional churn-analysis methods, which rely on simple segmentation or descriptive thresholds, often miss complex interactions among usage, revenue, and customer attributes. However, accurately predicting which customers will leave enables proactive, targeted retention strategies that safeguard profitability. We propose that supervised machine-learning models, properly engineered and tuned, can outperform naive heuristics in identifying at-risk business accounts (Chang et al., 2024; Sikri et al., 2024).

To test this, we used a real-world dataset of 8,453 Bulgarian enterprise customers, containing subscriber counts, revenue metrics, and customer segments; and built an end-to-end pipeline: data cleaning (imputing 19 % missing subscriber and ARPU values), feature engineering (ratios and ZIP regions), and dimensionality reduction via PCA (5 components, 97 % variance). We trained logistic regression (baseline and balanced), Random Forest, and Gradient Boosting classifiers, tuning ensembles with randomized search and evaluating on precision, recall, F1-score, and ROC AUC. The tuned Random Forest achieved a cross-validated ROC AUC

of **0.582** (test **0.567**) and raised churn recall to **21%** from **0%** at baseline, demonstrating the model's practical value in guiding retention campaigns.

II/ Methods

1) Dataset & Goal

We used the Mendeley Customer Churn Dataset (Tokmakov, 2024), which contains 8,453 anonymized records of business clients from a Bulgarian telecom operator 14 fields.

The goal of this project was to predict whether a business customer would churn:

- Binary churn indicator (CHURN: 1 = Yes, 0 = No).

Features using including: contract type, company size, service subscription patterns, activity ratios (e.g., active and suspended subscriber ratios), and revenue-related metrics.

a) Numeric features

1. Active_subscribers: Active lines on the account
2. Not_Active_subscribers: Unused assigned lines
3. Suspended_subscribers: Temporarily suspended lines
4. Total_SUBs: All lines (active + inactive + suspended)
5. AvgMobileRevenue: Mean monthly mobile revenue
6. AvgFIXRevenue: Mean monthly fixed-line revenue
7. TotalRevenue: Total monthly revenue (mobile + fixed).
8. ARPU (Average Revenue Per User): TotalRevenue divided by Active_subscribers.

b) Categorical features

1. CRM_PID_Value_Segment: CRM-assigned customer tier (e.g., Bronze, Silver, Gold)
2. EffectiveSegment: Business-size segment such as SOHO (Small Office/Home Office), VSE (Very Small Enterprise), SME.
3. KA_name: Key-account manager ID responsible for the client.
4. ZIP_region: First two digits of billing ZIP code

2) Data Cleaning

- Imputed missing subscriber counts (Not_Active_subscribers and Suspended_subscribers) to zero (19 % of rows).
- Recomputed missing ARPU = TotalRevenue/Active_subscribers, then filled remaining nulls with 0.
- Converted CHURN to integer; derived ZIP_region from billing ZIP.

3) Feature Engineering & PCA

- Engineered active_ratio and suspended_ratio
- Standardized numerics (zero mean, unit variance).
- One-hot encoded categoricals.
- Applied PCA on scaled numerics to retain 95% variance => 5 principal components (97.1% explained).

4) Model & Tuning

- Baselines: Logistic Regression; balanced Logistic (class_weight).
- Ensembles: Random Forest; Gradient Boosting.

- Hyperparameter Search: RandomizedSearchCV over 20 candidate RF configurations (n_estimators, max_depth, min_samples_split/leaf, max_features), 5-fold CV, optimizing ROC AUC.
- Evaluation Split: 80/20 stratified train/test to preserve churn proportion.

5) Evaluation Metrics

- Classification: precision, recall, F1-score (report per class).
- Ranking: ROC AUC.
- Interpretation: confusion matrices, permutation importance, and partial-dependence plots.

III/ Results

Table 1: Churn-Class Metrics & ROC AUC

Model	Precision	Recall	F1-score	ROC AUC
Logistic Regression (base)	0.00	0.00	0.00	0.580
Balanced Logistic Regression	0.08	0.53	0.14	0.580
Random Forest (baseline)	0.00	0.00	0.00	0.575
Tuned Random Forest	0.10	0.21	0.13	0.567
Gradient Boosting (baseline)	0.00	0.00	0.00	0.579
Logistic + PCA	0.08	0.52	0.14	0.572

Table 2: Overall Accuracy & F1-Score

Model	Accuracy	Macro F1	Weighted F1
Logistic Regression (base)	0.93	0.48	0.90
Balanced Logistic Regression	0.59	0.44	0.70
Random Forest (baseline)	0.93	0.48	0.90
Tuned Random Forest	0.82	0.52	0.85
Gradient Boosting (baseline)	0.93	0.48	0.90
Logistic + PCA	0.59	0.44	0.69

Table 3: Top 5 Features by Permutation Importance

Rank	Feature	Mean Importance	Std Dev
1	num__AvgMobileRevenue	0.0286	0.0173
2	num__TotalRevenue	0.023	0.0103
3	cat__ZIP_region_65	0.0071	0.0043
4	cat__KA_name_DI	0.0066	0.0039
5	num__ARPU	0.0064	0.0061

Key Findings: Tuned Random Forest is the best overall performer. Although the balanced logistic model reached higher recall (0.53), its precision (0.08) and F1 (0.14) remained low. The tuned RF (configured with 200 trees, max_depth=10, and min_samples_leaf=4) achieved: 21% churn recall vs 0% at baseline.

IV/ Discussion

The tuned Random Forest improved churn recall to 21% from 0%, a critical uplift for identifying at-risk customers. However, overall recall remains modest, reflecting the challenge of highly imbalanced data and subtle churn signals. The baseline logistic model, even when class-weighted, struggled to balance precision and recall, underscoring the necessity of non-linear ensembles.

1) Interpretation & Business Value:

- AUC near 0.57 indicates moderate discrimination but actionable risk stratification: the top 10 % risk bucket contains nearly half of churners, focusing retention resources effectively.
- Revenue-related features dominate importance rankings, suggesting that declines in mobile or total revenue are early churn indicators.

2) Hurdles & Resolutions:

- Class Imbalance: Initial models predicted no churn. We addressed this via class_weight = 'balanced' and tuning, but additional resampling (SMOTE) or cost-sensitive boosting may further improve recall.
- Data Quality: 19 % missing in subscriber fields required careful imputation to avoid bias.

3) Successes & Lessons:

- PCA reduced numeric dimensions from eight to five without degrading AUC, streamlining models where interpretability is less critical.
- Partial-dependence plots confirmed non-linear churn relationships (e.g., risk sharply rising when AvgMobileRevenue falls below a threshold), guiding targeted incentives.

4) Future Work:

- Segmented Models: Build separate models for high-value vs. low-value segments to capture distinct churn drivers.
- Ensemble Stacking: Combine RF and gradient boosting in a stacked classifier for incremental gains.
- Deployment Pipeline: Serialize the full preprocessing and model pipeline for real-time scoring and drift monitoring.

V/ Citations and Attributions

- **Dataset:** Tokmakov, D. (2024). *Customer Churn Dataset from a Bulgarian Business Telecom Operator*. Mendeley Data. https://data.mendeley.com/datasets/nrb55gr66h/1
- **Tools:** Python (pandas, scikit-learn, matplotlib, seaborn)
- **AI Assistance:** ChatGPT for report introduction, citation, structuring and grammar editing.

Chang, V., Hall, K., Xu, Q. A., Amao, F. O., Ganatra, M. A., & Benson, V. (2024). Prediction of customer churn behavior in the telecommunication industry using machine learning models. *Algorithms, 17*(6), 231.

https://doi.org/10.3390/a17060231

Sikri, A., Jameel, R., Idrees, S. M., & Kaur, H. (2024). Enhancing customer retention in telecom industry with machine learning driven churn prediction. *Scientific Reports, 14*, Article 13097.

https://doi.org/10.1038/s41598-024-63750-0

VI/ Figures and Captions

Figure 1: Correlation matrix showing relationships among subscriber activity and revenue metrics.

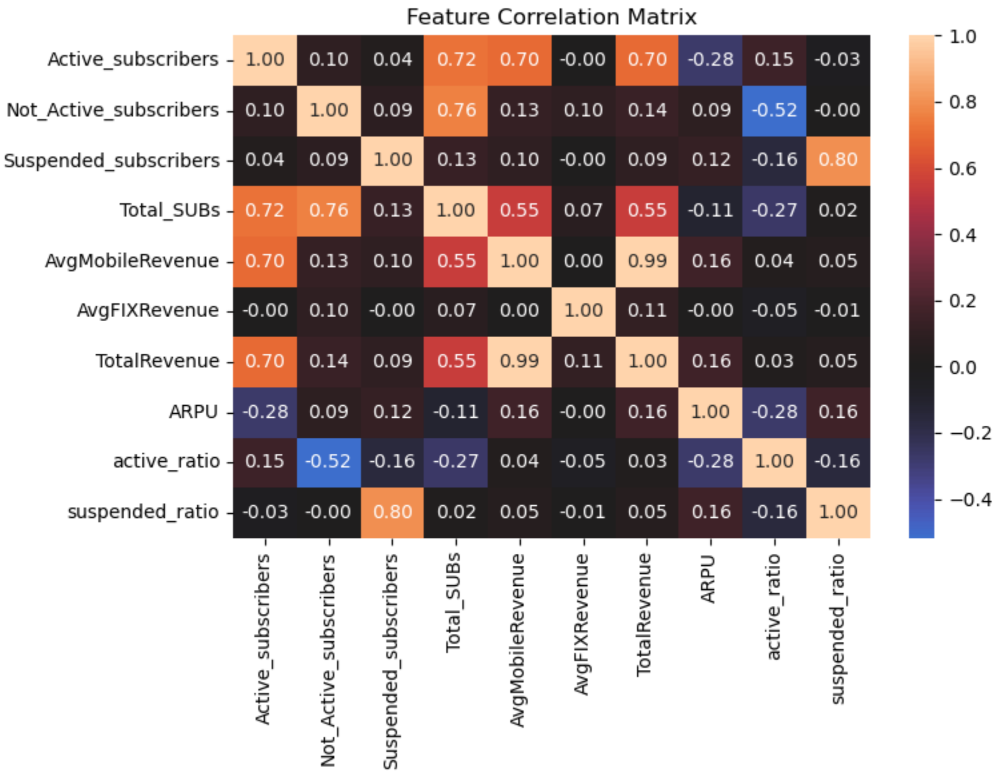


Figure 2: Top 10 feature importances based on permutation importance analysis.

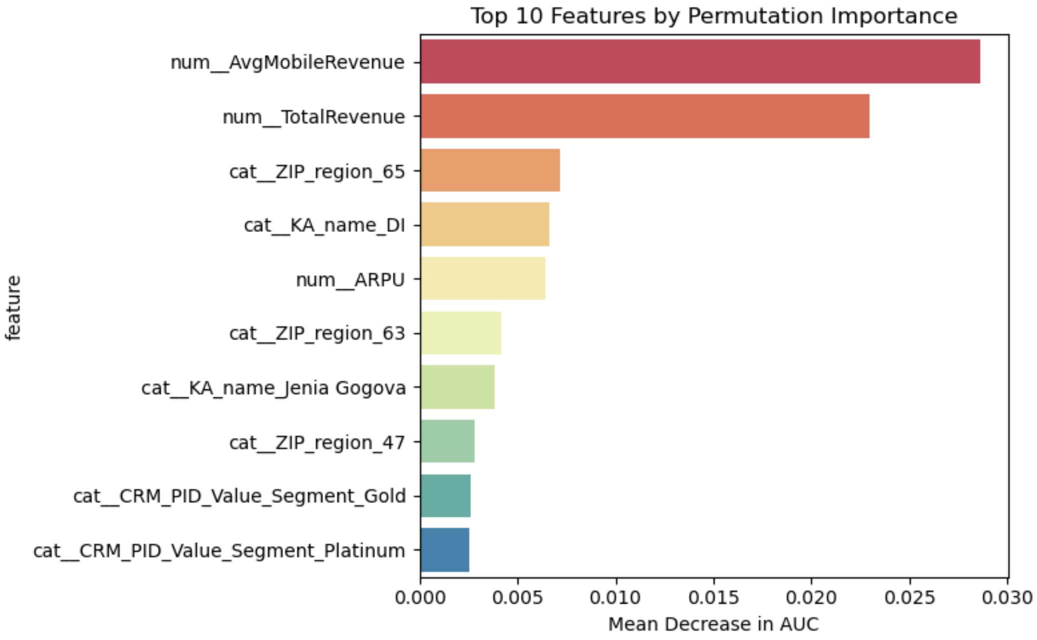


Figure 3: ROC curve and Precision-Recall curve for the Tuned Random Forest model.

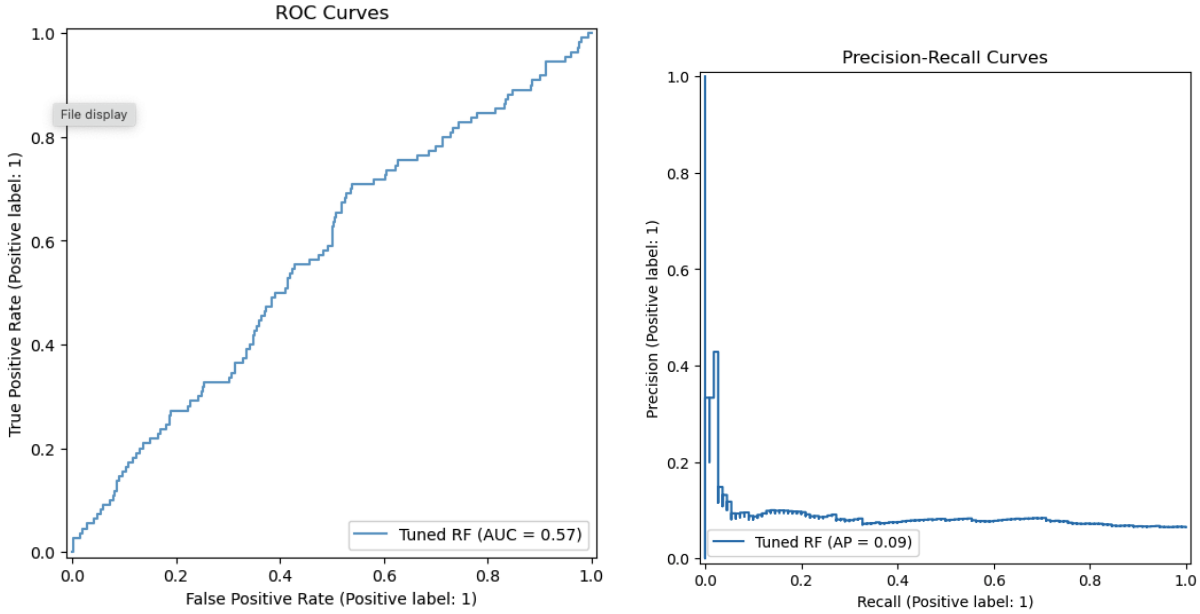


Figure 4: Active Subscriber Ratio and Suspended Subscriber Ratio distributions by churn status.

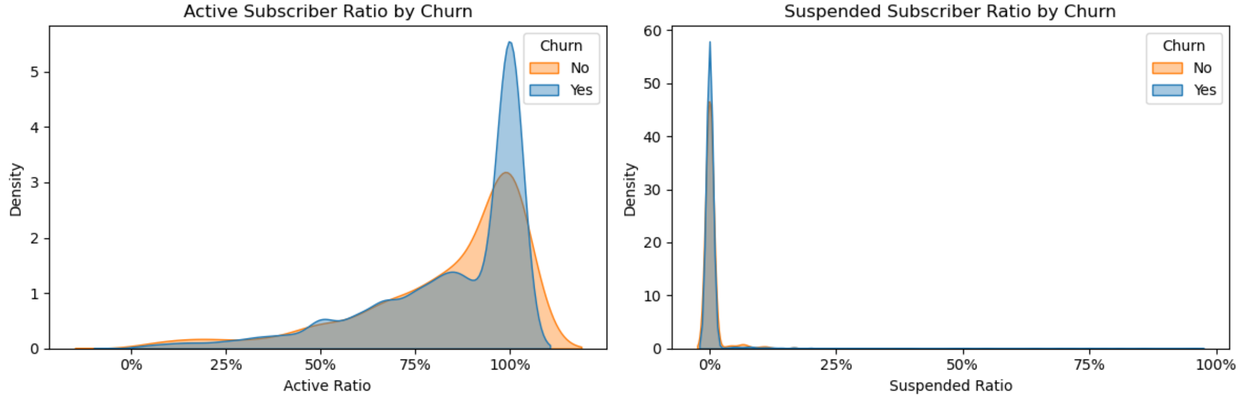


Figure 5: ARPU and Total Revenue distributions segmented by churn.

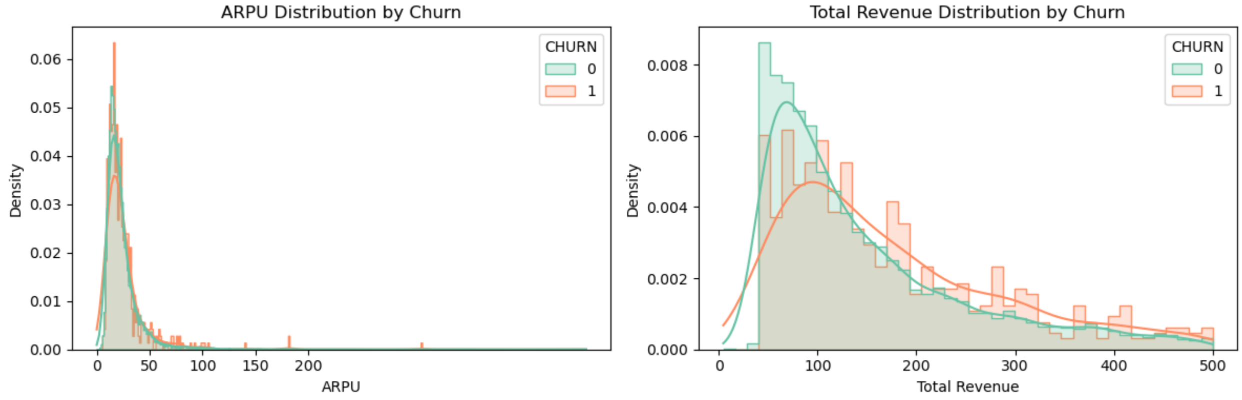


Figure 6: Churn Rate by CRM Segment & Effective Segment

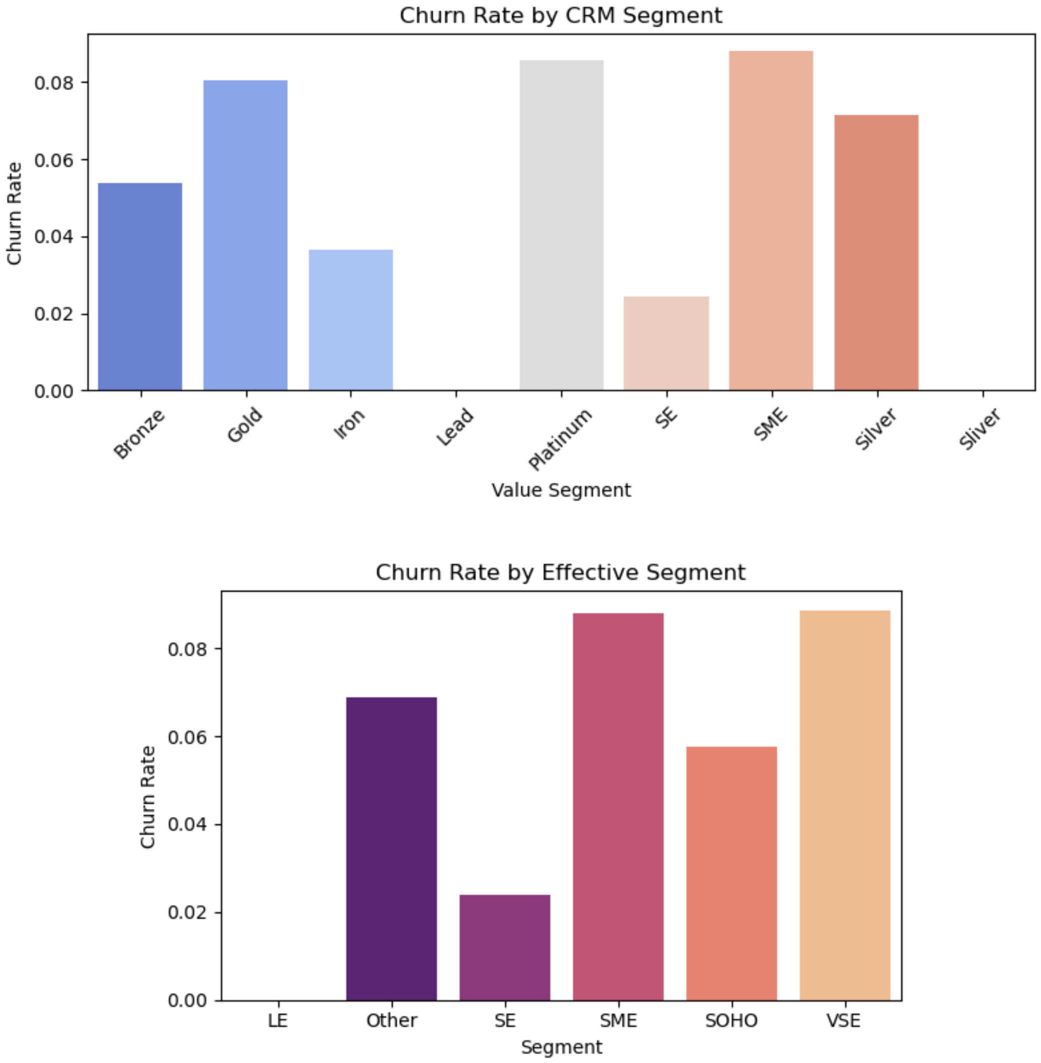
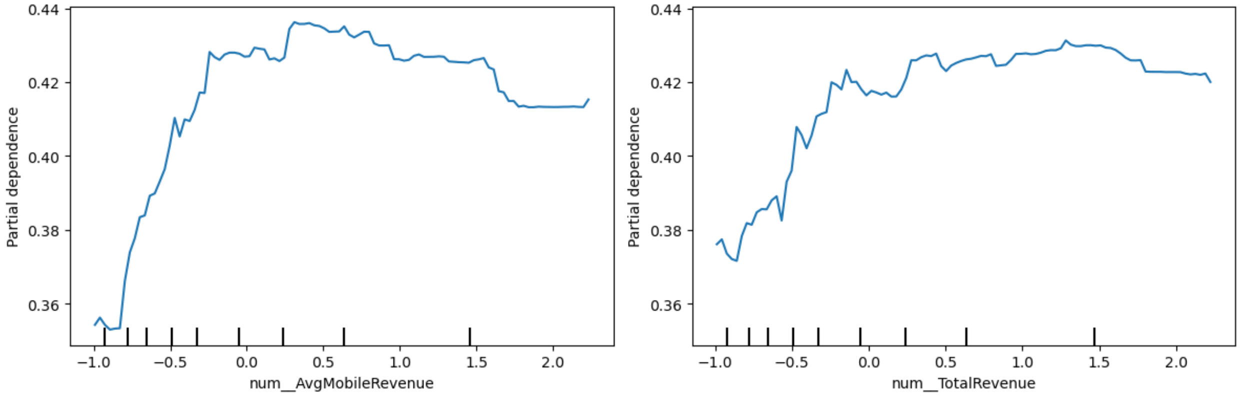


Figure 7: Partial dependence top features


Partial dependence
num__AvgMobileRevenue
Partial dependence
num__TotalRevenue