Comprehensive Classification Project Report

1. Objective of the Analysis

The primary goal of this project was to build, compare, and evaluate various classification models to predict customer churn in the telecom industry1. Retaining customers is more cost-effective than acquiring new ones, making churn prediction a critical task2. The report not only evaluates model performance but also emphasizes the importance of addressing class imbalance in the dataset3. The models evaluated were Logistic Regression, Support Vector Machines (SVM), and Random Forest Classifiers.

---

2. Dataset Description

The dataset contains information on telecom customers, including demographics, account details, and subscription services5. The target variable,

Churn, indicates if a customer has left the service6666. The dataset is moderately imbalanced, with approximately 73% of customers being non-churners and 27% being churners7. This imbalance could lead to potential bias in models that don't account for it8. The features include both categorical variables (e.g., gender, contract type) and numerical variables (e.g., tenure, monthly charges)9.

---

3. Data Preparation

Before model training, several preprocessing steps were performed10:

- Unique identifiers like `customerID` were removed11.
- Missing values, particularly in the `TotalCharges` feature, were handled by coercing the data to a numeric type and imputing where necessary12.
- Categorical variables were encoded using binary mapping or one-hot encoding13.
- Numerical features were standardized to ensure they contribute fairly to models14.
- Class imbalance was addressed using two strategies: applying **class weights** within algorithms and oversampling the minority class using **SMOTE** (Synthetic Minority Over-sampling Technique)15.

---

4. Model Training & Evaluation

Multiple classifiers were trained and evaluated under different imbalance-handling strategies16. The performance of each model was summarized in a table showing key metrics like accuracy, precision, recall, and F1-score for the "Churn" class17.

Model	Imbalance Handling	Accuracy	Precision (Churn)	Recall (Churn)	F1-score	Remarks
Logistic Regression	Balanced Weights	0.759	0.535	0.719	0.613	Strong recall, interpretable coefficients
Logistic Regression	SMOTE	0.756	0.530	0.717	0.609	Similar to weights, but adds pipeline complexity
SVM	None	0.780	0.560	0.520	0.540	Good accuracy, but under-detects churners
SVM	Balanced Weights	0.740	0.500	0.710	0.590	Significant recall boost, lower precision
Random Forest	None	0.800	0.680	0.500	0.580	High accuracy, biased toward majority class
Random Forest	Balanced Weights	0.780	0.620	0.690	0.650	Balanced trade-off, strong F1-score

---

5. Final Model Recommendation

Based on the evaluations, the following recommendations are made18:

- **For interpretability: Logistic Regression with class weights** is the preferred model, as its coefficients directly reveal churn drivers19.
- **For a balance of precision and recall: Random Forest with class weights** offers the best trade-off, making it ideal for deployment where both accuracy and fairness are

important20. * **For maximizing recall: SVM with class weights** significantly improves the detection of churners, which may be useful if the cost of missing a churner is high21.

---

6. Key Insights

- Handling class imbalance is crucial, as models that don't address it tend to underpredict churners22.
- SVM shows high recall potential but at the cost of precision, which is suitable for aggressive customer retention strategies23.
- Random Forest with balanced handling provides the most consistent results, effectively capturing both churners and non-churners24.
- Logistic Regression is valuable for its simplicity, interpretability, and strong baseline recall performance25.

---

7. Next Steps

To further enhance the project, the following steps could be taken26:

- **Threshold optimization:** Adjust the probability cutoffs to align with the business costs of false positives versus false negatives27.
- **Hybrid approach:** Deploy the Random Forest model for real-time churn detection, while using the Logistic Regression coefficients to create explanatory dashboards for business teams28.
- **Feature engineering:** Create new behavioral features from service usage or complaint history to improve predictive power29.
- **Advanced modeling:** Explore more powerful algorithms like Gradient Boosting and XGBoost30.
- **Deployment:** Integrate the final model into CRM systems for real-time predictions to enable proactive retention campaigns31.