IS 804 Report: Predicting Credit Card Customer Churn Based on Income Level, Demographic, and Behavioral Features

Marjory Pineda

May 2025

1 Introduction

The literature reveals that customer retention is a central challenge for companies providing financial products and services to customers. Acquiring new customers, as well as retaining existing customers, often comes with a cost, and understanding the factors that contribute to customers leaving is a key focus for companies. The phenomenon in which customers discontinue their relationship with a particular company is known as customer churn. Churn can be voluntary, meaning customers actively leave, or involuntary due to inactivity with a company's products or services, or company decisions. Whether voluntary or involuntary, customer churn directly impacts a company's profitability and long-term growth [4]. For example, in highly competitive markets, retaining existing customers is often more cost-effective than acquiring new ones [1], which makes churn prediction a critical area of research for companies and their customer relationship management practices. In the context of predictive modeling for customer churn, that is, the likelihood that a customer will discontinue their relationship with a company, quantitative analysis methods, including machine learning and statistical modeling, offer data-driven tools to proactively identify customers who are at-risk of leaving or switching. These data-driven tools can help companies implement strategies for customer retention and satisfaction. While previous work has focused on traditional models that often rely on transactional or behavioral data, demographic variables such as income level remain underexplored as primary predictors [8, 10]. Understanding the role of income level in customer disengagement from companies can help identify patterns of market segmentation and financial variability.

Therefore, this project aims to explore an under-analyzed area of churn prediction — the predictive value of income level. The study investigates the following interrelated research questions within the context of banking, more specifically, predicting customer churn in credit card usage:

- **RQ1: Can we accurately predict whether a customer will churn based solely on their income level?**

- **RQ2: How does the inclusion of *behavioral* and *demographic features* alongside income improve the accuracy and interpretability of customer churn prediction models?**

By isolating income as a predictive feature and later integrating it with a broader set of variables, this study contributes to the growing literature on equitable and interpretable churn modeling. The findings have practical implications for financial institutions, including banks, seeking to improve customer retention strategies through targeted, data-driven insights.

2 Related Works

2.1 Customer Churn and Predictive Analysis

Customer churn prediction has been extensively studied across multiple industries, including telecommunications, banking, insurance, and retail. In the telecommunications sector, for instance, Huang et al. [5] demonstrated the effectiveness of classification algorithms such as decision trees and neural networks in predicting churn. Furthermore, Lalwani et al. [9] developed a comprehensive machine learning approach that applies Random Forest and Gradient Boosting to enhance prediction accuracy. In the banking sector, Dana Al-Najjar et al. [1] employed logistic regression and decision trees to predict credit card churn, identifying transaction patterns as critical features in the prediction of customer churning. Similarly, Prabadevi et al. [12] used support vector machines and k-nearest neighbors for churn analysis, showing that ensemble techniques generally outperformed individual models. Finally, De and Prabu [4] provide a systematic literature review highlighting a wide range of machine learning approaches and their effectiveness across different sectors, emphasizing an increasing trend toward hybrid models that combine predictive accuracy with interpretability. Furthermore, comparative studies, such as those by Özer Çelik et al. [3], explore the strengths and weaknesses of various modeling techniques, highlighting the trade-offs between performance metrics and operational deployment. These works collectively demonstrate that while behavioral metrics dominate churn prediction models across various sectors, including banking, demographic and contextual variables remain less explored in the literature.

2.2 Income and Consumer Behavior

An individual's income level is an important variable in consumer behavior, influencing financial habits, risk tolerance, and patterns of how they use a particular product, service, or company. Kamakura et al [7] argue that income segmentation significantly affects market positioning and response to marketing interventions. Additionally, Baker et al. [2] detail how income volatility contributes to customer decision-making under uncertainty, which can furthermore exacerbate churn likelihood in lower-income levels. Finally, Kaya et al. [8]

suggest that behavioral attributes, when filtered through the lens of financial well-being, can better predict churn in economically vulnerable populations.

2.3 Income-Based Segmentation in Predictive Analysis

Despite income being a well-known stratifier in consumer research, its direct role in churn modeling has often been secondary to transactional or engagement features, especially as it relates to the banking sector. Studies such as Larivière & Van den Poel [11] and Verbraken et al [13] have used income as a supporting feature only and did not delve into the contributions of income, or in other words, did not isolate its predictive power. However, income-based segmentation has been employed successfully in other sections, such as marketing analytics [7] and is increasingly recognized in the literature on churn [8] for its potential to capture customer heterogeneity.

This project seeks to build on the foundation presented in the related works section by explicitly focusing on and modeling the impact of income on churn likelihood in credit card usage and evaluating how much predictive value it adds when combined with other features such as behavioral and demographic.

3 Data

This project will utilize data from the **Bank Churners dataset** [6], which is publicly available on Kaggle, a platform primarily used for data science competitions. The Bank Churners dataset consists of **10,127 anonymized customer records on credit card usage**. Each record represents an individual who owns a credit card account. The dataset also includes **23 features** capturing a customer's demographic, behavioral, and financial information, along with a **binary target variabl**e indicating **customer churn status.**

3.1 Key Variables

The dataset is comprised of `Attrition_flag`, which is the **target variable** (values include existing customer, decoded as 0, and attrited customer, decoded as 1). The **demographic features** include gender, education level, marital status, income category (less than$40K, $40K–$60K, $60K–$80K, $80K–$120K, $120K+, and Unknown), and card category(Blue, Silver, Gold, Platinum). Finally, the **behavioral account features** include customer age, months the account has been active, total number of products held by customer, months inactive in the last 12 months, contacts count in the last 12 months, credit limit, total revolving balance, average open to buy, total transaction amount in the last 12 months, total number of transactions in the last 12 months, and average credit utilization ration.

3.2 Data Characteristics and Limitations

- **Categorical variables** (e.g., `Income_Category`, `Marital_Status`) will be encoded using one-hot encoding or ordinal schemes where appropriate.
- **Missing values**: Some categorical features contain "Unknown" as a category; this will be carefully handled during preprocessing.
- **Class imbalance**: The dataset is somewhat imbalanced, with the majority of customers labeled as "Existing Customer." This will be addressed using stratified sampling or resampling techniques if necessary.

3.3 Relevance for Course Project

The dataset is well-suited for binary classification tasks and supports a wide range of modeling approaches covered in the IS 804 course, including logistic regression, support vector machines, decision trees, and additive models. The dataset's inclusion of demographic, behavioral, and **income-related features** makes it relevant to the research questions focused on income-based churn prediction.

4 Methods

The analysis of the Bank Churners dataset was carried out in the following steps: 1. data preprocessing, 2. baseline modeling for RQ1, and 3. full-feature modeling for RQ2. Detailed procedures are described below.

4.1 Data Preprocessing

Before analyzing the dataset, a series of preprocessing steps were performed to address the first research question (RQ1), which investigates whether credit card customer churn can be predicted using income level alone. These steps included cleaning the data, encoding categorical variables, and creating training and testing data splits using stratified sampling.

After loading the dataset, a unique identifier column (`CLIENTNUM`) was removed as it did not contain predictive information. A new binary target variable (`Churn`) was created by encoding the `Attrition_Flag` column. Customers labeled as *Attrited Customers* were assigned a value of 1 (churn), and *Existing Customers* were assigned a value of 0 (not churn).

For RQ1, the only predictor used was `Income_Category`. This variable, originally stored as a string with categories (e.g., "Less than $40K", "$40K–$60K", "Unknown"), was encoded into numerical values using Label Encoding. The encoded values were stored in a new column, `Income_Category_Encoded`. The data was then split into training and testing sets using an 80/20 stratified split to preserve class proportions (`random_state=42`).

4.2 Selected Models

To model customer churn—a binary classification problem—three supervised learning algorithms were selected based on their theoretical differences and relevance to classification tasks:

- **Logistic Regression**, which models the log-odds of the binary outcome using a linear function of the input features.
- **Decision Tree Classifier**, which partitions the feature space using a series of if-then conditions, capturing nonlinear patterns and interactions.
- **Support Vector Machine (SVM)**, which seeks the optimal boundary between classes by maximizing the margin, offering robustness in low- and high-dimensional feature spaces.

These models were chosen to provide diverse modeling perspectives: parametric (logistic), rule-based (tree), and margin-based (SVM). All models were evaluated using precision, recall, F1-score, and ROC AUC. ROC AUC was emphasized due to class imbalance in the dataset, with churned customers comprising approximately 16% of the total population.

4.3 RQ1 – Modeling using Income Only

For RQ1, models were trained using only the `Income_Category_Encoded` feature. Each model was trained on the training set (80%) and evaluated on the test set (20%). The goal was to determine whether income category alone provides enough signal to predict churn accurately.

4.4 RQ2 – Modeling using All Behavioral and Demographic Features

For RQ2, the same three classifiers were retrained using a broader feature set that included demographic variables (e.g., `Gender`, `Marital_Status`, `Education_Level`), behavioral metrics (e.g., `Total_Trans_Ct`, `Contacts_Count_12_mon`), and financial indicators (e.g., `Credit_Limit`, `Avg_Utilization_Ratio`). Categorical features were one-hot encoded and numeric features were standardized. Models were again trained on the 80% training set and evaluated on the 20% test set.

5 Results

5.1 Data Preprocessing Results

The dataset contained 10,127 customer records, of which 1,627 (16.1%) represented churned customers. After encoding the `Income_Category` variable, five distinct encoded categories were used in the modeling. The resulting training set contained 8,101 observations and the test set 2,026 observations, with a single predictor used for RQ1.

5.2 RQ1 – Modeling Results (Income Only)

All three models performed poorly when using only income as a predictor. ROC AUC scores hovered slightly above 0.5, indicating that **income alone is not a strong predictor of churn**. While income may serve as a relevant demographic marker, it does not provide enough discriminatory power when isolated. These results support the move toward richer, multi-variable modeling explored in RQ2 (see Figure 1).

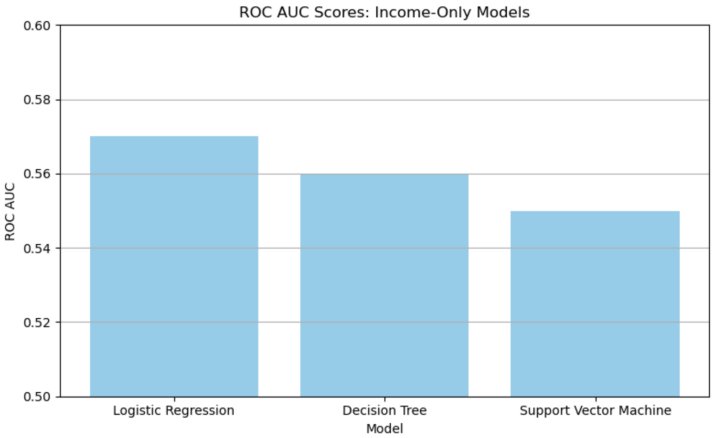

Figure 1: ROC AUC Scores: Income-Only Models

5.3 RQ2 – Modeling Results (All Features)

The results of the full-feature modeling revealed that incorporating demographic and behavioral variables substantially improved predictive performance compared to income-only models. All three classifiers—logistic regression, decision tree, and SVM—achieved high ROC AUC scores (0.90–0.94), indicating excellent discriminatory capability (see Figure 2).

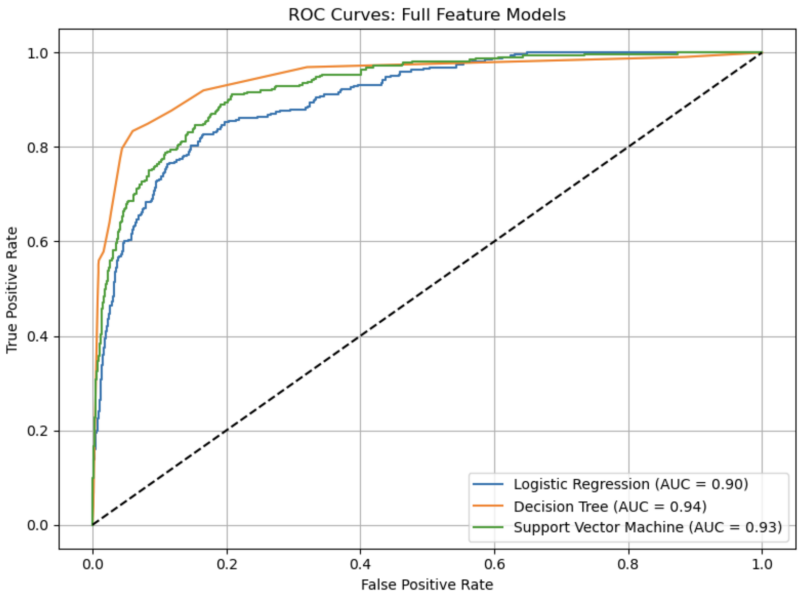

Figure 2: ROC Curves: Full Feature Models

Given that the ROC AUC scores were significantly high, a confusion matrix analysis was included to better understand how each model behaved across churn and non-churn classes. This was especially important given the class imbalance in the dataset, where churned customers represent only about 16% of observations. While ROC AUC and accuracy summarize overall model performance, the confusion matrix helps to reveal the distribution of true positives, false positives, true negatives, and false negatives, making it possible to evaluate each model's trade-offs between precision and recall. For example, a model that achieves high accuracy may still fail to identify churners if its recall is low.

The decision tree model achieved the best balance of precision (0.93) and recall (0.93) for identifying churned customers, making it the most effective in minimizing both false positives and false negatives. In contrast, logistic regression and SVM demonstrated slightly lower precision (0.88 and 0.90, respectively) and slightly lower recall (0.89 and 0.90, respectively).

These results reinforce the value of behavioral and demographic variables in churn prediction and highlight the importance of model selection based on business goals, whether minimizing false positives (e.g., costly retention offers to likely non-churners), or maximizing true churn capture. Including the confusion matrix in the evaluation process supports this kind of nuanced, business-aligned model interpretation.

6 Discussion

This study examined the extent to which customer churn can be predicted using income level alone (RQ1), and how predictive performance improves when

behavioral and demographic features are incorporated (RQ2). Using the Bank Churners dataset, we employed three supervised classification models—logistic regression, decision trees, and support vector machines—commonly applied in churn analysis literature [1, 12, 3].

The approach in this study followed a two-phase modeling strategy, beginning with a minimal predictive setup (income only), followed by an expanded feature set approach aligned with best practices in churn prediction [5, 9, 4]. Results from RQ1 demonstrated that income category alone does not provide sufficient predictive power, with all models yielding ROC AUC scores slightly above 0.5. This aligns with prior findings suggesting that while income segmentation may correlate with customer behavior [7, 8], it is rarely a strong standalone predictor of churn.

In contrast, the results from RQ2 showed significant improvements. By integrating behavioral indicators such as transaction volume, inactivity duration, and utilization ratios—variables that have shown high predictive value in the literature [5, 8], all models achieved ROC AUC scores above 0.90. The decision tree classifier achieved the best balance of precision and recall, while logistic regression and SVM offered slightly more conservative churn predictions. These findings reinforce that combining behavioral and demographic features yields far more effective models, echoing conclusions from earlier work on hybrid predictive approaches [4, 9, 10]. The approach in this study included confusion matrix analysis to better understand the trade-offs between false positives and false negatives which is a critical consideration in real-world churn management where organizational outreach efforts are often costly. As emphasized in the literature, the strategic choice of model should depend on business goals: reducing unnecessary retention efforts (favoring high precision), or maximizing detection of at-risk customers (favoring high recall)[**?**, **?**]. The main contribution of this report is a structured, comparative analysis of churn prediction using income alone versus a broader socio-behavioral feature set, applied to a real-world dataset. By isolating income as a predictor and systematically incorporating other features, this report provides empirical support for the claim that income, while contextually relevant, is not a sufficient standalone predictor of churn. This reinforces the importance of behavioral and interactional data in churn modeling and provides a baseline for future segmentation-driven approaches in customer analytics.

6.1 Limitations and Future Work

Several limitations should be noted. First, the dataset used is from a single institution and may not generalize to other financial contexts or customer bases. Second, the income feature was categorical and coarse-grained, which limited the ability to model nuanced economic influences on churn. Future research could explore the use of continuous or externally sourced income estimates and investigate how economic shocks or volatility affect churn likelihood[2]. Additionally, while I tested three interpretable models, other machine learning methods such as ensemble learning (e.g., XGBoost or Random Forests) could

offer deeper insights into feature importance and model transparency.

References

[1] Al-Najjar, D., Al-Rousan, N., and Al-Najjar, H. Machine learning to develop credit card customer churn prediction. *Journal of Theoretical and applied electronic commerce research 17*, 4 (2022), 1529–1542.

[2] Baker, S. R., Baugh, B., and Sammon, M. C. Measuring customer churn and interconnectedness. Tech. rep., National Bureau of Economic Research, 2020.

[3] Çelik, O., and Osmanoglu, U. O. Comparing to techniques used in customer churn analysis. *Journal of Multidisciplinary Developments 4*, 1 (2019), 30–38.

[4] De, S., and Prabu, P. Predicting customer churn: A systematic literature review. *Journal of Discrete Mathematical Sciences and Cryptography 25*, 7 (2022), 1965–1985.

[5] Huang, B., Kechadi, M. T., and Buckley, B. Customer churn prediction in telecommunications. *Expert Systems with Applications 39*, 1 (2012), 1414–1425.

[6] Kaggle. Bankchurners.csv, accessed 2025.

[7] Kamakura, W., Mela, C. F., Ansari, A., Bodapati, A., Fader, P., Iyengar, R., Naik, P., Neslin, S., Sun, B., Verhoef, P. C., et al. Choice models and customer relationship management. *Marketing letters 16* (2005), 279–291.

[8] Kaya, E., Dong, X., Suhara, Y., Balcisoy, S., Bozkaya, B., et al. Behavioral attributes and financial churn prediction. *EPJ Data Science 7*, 1 (2018), 41.

[9] Lalwani, P., Mishra, M. K., Chadha, J. S., and Sethi, P. Customer churn prediction system: a machine learning approach. *Computing 104*, 2 (2022), 271–294.

[10] Lappeman, J., Franco, M., Warner, V., and Sierra-Rubia, L. What social media sentiment tells us about why customers churn. *Journal of Consumer Marketing 39*, 5 (2022), 385–403.

[11] Larivière, B., and Van den Poel, D. Predicting customer retention and profitability by using random forests and regression forests techniques. *Expert systems with applications 29*, 2 (2005), 472–484.

[12] Prabadevi, B., Shalini, R., and Kavitha, B. R. Customer churning analysis using machine learning algorithms. *International Journal of Intelligent Networks 4* (2023), 145–154.

[13] Verbraken, T., Verbeke, W., and Baesens, B. Profit optimizing customer churn prediction with bayesian network classifiers. *Intelligent Data Analysis 18*, 1 (2014), 3–24.