Predicting and Preventing Customer Churn Using Both Classical and Robust Techniques in Machine Learning

Felicia Nappi, Jaeyoon Wang

15.095 Fall 2024

Intro/Problem Significance

Customer churn is a critical issue for businesses, with rates reaching 30%[1] in telecommunications and 10%[2] in consumer goods industries, exceeding the recommended benchmark of 5%. These high churn rates result in significant revenue losses and operational inefficiencies, emphasizing the importance of customer attainment. Churn is driven by subtle perturbations in customer behavior influenced by service quality, pricing, and external factors, making it challenging to predict and prevent. Identifying common features among customers who churn versus those who stay loyal allows companies to target at-risk groups with tailored interventions, minimizing churn and enhancing retention efforts.

This paper compares the predictive power of classical logistic regression and k-means clustering with their robust counterparts, analyzing two datasets from distinct domains: telecommunications (Telco dataset) and consumer goods (iFood dataset). Robust models incorporate uncertainty, addressing noise and variability in customer data, but their value in clean datasets remains unclear. By evaluating the performance and trade-offs of these approaches, this project identifies conditions where robust methods enhance prediction accuracy and highlights actionable insights for customer retention strategies. Our findings aim to bridge the gap between theoretical robustness and practical application in churn prevention.

Data

To address the aspects of predicting and preventing customer churn, two separate datasets were used, both of which are from Kaggle. The predicting dataset[3] originated on IBM regarding customers from a telecommunications company in California from Q3 2019. The covariates included demographic information about the customer and their family, the type of services the company provided to the customer, and how the customer paid for their services (paperless billing, credit card statements, etc.). The response variable is whether or not the customer chose to churn during the fiscal quarter selected.

---

[1] https://www.akkio.com/post/telecom-customer-churn
[2] https://customergauge.com/blog/average-churn-rate-by-industry
[3] https://www.kaggle.com/datasets/blastchar/telco-customer-churn

The preventing dataset[4] was adapted from a Brazillian food delivery service, called iFood. The dataset covariates also included customer demographics but also included how much customers spent on different types of food products, and how they chose to pay for the products (in the store, online, etc.).

While a lot of feature engineering and data manipulation did not have to occur due to the datasets being readily available, more sophisticated/theoretical methods were chosen to make sure this was a sufficient project for the scope of the course. The variables MonthlyCharges and TotalCharges were dropped from the Teleco dataset, as they were creating multicollinearity with other variables, which was confirmed by both a collinearity matrix and VIF values. No variables were dropped from the iFood.

Methods

Logistic Regression (Classical and Robust)

Classical logistic regression was the first model used, which can be used as a baseline when compared to more complex methods. The baseline formula is as follows:

$$P(Y = 1) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\ldots+\beta_p x_p)}}$$

Where β_0 is the intercept and $\beta_1 x_1 + \ldots + \beta_p x_p$ is the cross product of the p features and the corresponding coefficients.

However, it can be argued that some of these variables mentioned in the dataset can be perturbed, especially variables such as the internet as it can be very unstable for customers. Robust logistic regression is then implemented, with an uncertainty set in features. Uncertainty in features was chosen as they seemed better in the context of this issue as opposed to labels since consumer information can be incorrectly entered or the human behavior aspect can change the response of a customer instantly. With the uncertainty set in features defined as:

$$U_x = \{\Delta X \epsilon R^{n\,x\,p} \mid \left|\left|\Delta x_i\right|\right| \leq \rho,\ i = 1,\ldots,n\}$$

[4] https://www.kaggle.com/datasets/jackdaoud/marketing-data?select=ifood_df.csv

The uncertainty set and with robust counterpart done in the course is defined as:

$$max_{\beta,\beta_o} - \sum_{i=1}^{n} log(1 + e^{-y_i(\beta^T x_i + \beta_0) + \rho||\beta||_{q^*}})$$

Where ρ is the regularization parameter, and L_{q^*} is the dual normal of L_q.

For the norm used, the L1 norm was chosen. While L2 may be "smoother" in closed form, for the context of this problem, feature selection is wanted, which is where the L1 norm becomes extremely useful. Cross-validation was used to find the appropriate ρ for the strength of regularization, and 0.01 was selected as the most appropriate for maximizing performance (in this context, the Maximum Likelihood Estimator with the logistic loss function).

K-Means (Classical, Interpretable Clustering, and Robust K-Means)

Another way to appropriately visualize churn versus non-churn customers is with distinct clusters to see if they have similar features that make them choose to churn, versus retain services. One of the most common clustering methods in practice is K-Means, where observations are partitioned into K clusters.

The original k-means algorithm can formulated as this optimization problem:

$$min_{C_1,\dots,C_k} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C^k} \sum_{j=1}^{p} (x_{ij} - x_{i'j}) \right\}$$

Where p is the dimensionality of the data points, c_k is the number of points in k clusters, and x_{ij} is the jth feature of the ith data point.

It can be argued that adding robustness to this algorithm will not scale as well as in robust logistic regression. The method was chosen to maintain relevance to the course (machine learning focusing on optimization problems). The paper Robust k-means: a Theoretical Revisit paper by Alexandros

Georgogiannis[5] was used as inspiration for implementation. The entirety of the paper was not implemented due to time constraints, but the basic ideas were implemented, including:

$$c_k^{new} = \frac{1}{|X_k|}\Sigma_{x_i \varepsilon X_k} x_i$$ (calculating the centroid using the mean of the assigned point)

$$d_{max} = max_{x_i \epsilon X_k} \left|\left| x_i - c_k^{new} \right|\right|_2$$ (calculating the max distance between any given point and the centroid calculated)

$$c_k^{adjusted} = c_k^{old} + \lambda \frac{c_k^{new} - c_k^{old}}{d_{max}}$$ (if the max distance goes above the threshold, we adjust the centroid)

A lambda/threshold (λ) of 1 was chosen, after cross-validation of multiple values for the parameter.

Results

Logistic Regression

The Telco dataset performed a test accuracy of 80.8% on classical logistic regression, and it showed to have significance in the β at the 0.1 level in:

- Length of subscription (in months)
- The type of contract the subscriber had
- If they had fiber-optic internet
- If the subscriber had multiple lines
- If they stream movies

Once the robust regression was implemented with the parameters discussed above, the test accuracy dropped slightly to 78.0%. While this is not ideal with what this paper is proposing, it is not such a significant drop to cause concern.

---

5 https://proceedings.neurips.cc/paper_files/paper/2016/file/80a8155eb153025ea1d513d0b2c4b675-Paper.pdf

The iFood dataset had a test accuracy of 86.7% in the baseline model, with significance in the following covariates at the 0.1 level:

- Method of purchasing: in catalog, deals, store, online
- Recency of purchase
- If a household has a teen in the home
- Number of days as a customer

Implementing robustness led to a sharp decline in test accuracy, dropping to 49.9% with the robust methodology. Such a notable drop in accuracy shows that robustness may not be the best methodology in a logistic setting for this dataset.

Clustering

With the Telco dataset, the Silhouette score increased significantly from 0.221 in the classical model to 0.734 in the robust model. This improvement suggests that Robust K-Means provided more well-defined clusters for defining churn versus non-churn consumers. The model is potentially handling outliers or noisy data more effectively, which results in more compact and well-separated clusters. The increased Silhouette score indicates that the clusters were more cohesive internally and better separated.

For the iFood dataset, however, the Silhouette score decreased slightly from 0.891 to 0.849 with Robust K-Means. This suggests that the original clusters formed using classical K-Means were already well-separated, and the introduction of robustness might have marginally reduced the clarity of the boundaries between clusters. Given the high initial Silhouette score, it is possible that the dataset did not contain enough noise to benefit from the robust methodology, which may explain why Robust K-Means underperformed slightly compared to the classical version. The Silhouette score here effectively captures the quality of clustering by measuring both cohesion (similarity within clusters) and separation (difference between clusters).

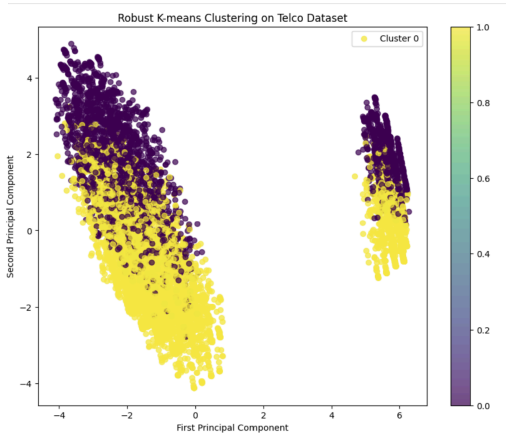

Figure 1. Teleco Robust K-Means Visualization

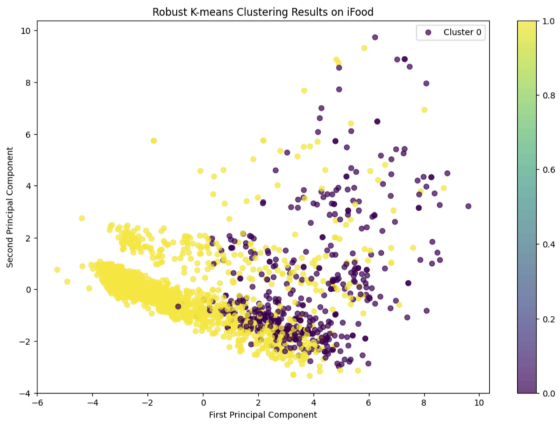

Figure 2. iFood Robust K-Means Visualization

Discussion/Conclusion

While both datasets used different features, they emphasized the importance of customer loyalty. Statistical significance was emphasized in both datasets, which shows a consistency in the importance of how long customers have been using the company's services and/or products. Based on the findings of the paper, companies should perhaps focus on keeping their long-term customers and maintaining that loyalty, as opposed to constantly trying to bring in new consumers and not being able to provide services as efficiently.

It can be hypothesized that the contrasting effects of robust methodologies on the two datasets can be attributed to differences in the data quality and structure. The Telco dataset likely benefited from handling noisy points or outliers, while the iFood dataset, being cleaner and more structured, did not require additional robustness, leading to a minor decline in cluster quality. The Telco dataset had several categories for their categorial data (for example, Internet Service had 3 categories and payment method had 4 categories), while the covariates of the iFood dataset were all either numerical or binary. The datasets were also smaller in nature than "big data" with the Telco and iFood datasets having approximately 7000 and 2200 observations, respectively.

It was seen in this case that robustness did not help with test accuracy, and in fact, significantly decreased performance with the iFood dataset. Robustness is important for noise though, and maybe the datasets did not contain enough noise as they are pre-cleaned datasets from Kaggle. Perhaps an extension of this could be to test the models on marketing data scraped from the internet from a platform such as Google Analytics. Robustness can sometimes work well for unbalanced data, as models can sometimes favor the majority class. This was the case here as both datasets have around 20 percent of observations churning/not accepting the campaign, while 80 percent of the observations are non-churn. However, there is also the need to do proper sampling and with the size of the datasets, it may not be applicable in this scenario.

If there was more time to expand on this topic, something feasible could have been the implementation of the entirety of Georgogiannis's paper, as there are several propositions and theorems that could change the performance of the algorithm. Another interesting expansion would be comparing Georgogiannis's paper with the Robust Classification paper published by Birstimias in 2019[6] (where the robust logistic regression used at the beginning of this paper came from), and seeing the similarities and differences in their formulations and ideas.

[6] https://doi.org/10.1287/ijoo.2018.0001

Works Cited

Bertsimas, D., Dunn, J., Pawlowski, C., & Zhuo, Y. D. (2019). Robust classification. *INFORMS Journal on Optimization*, *1*(1), 2–34. https://doi.org/10.1287/ijoo.2018.0001

Georgogiannis, A. (2016). *Robust k-means: A theoretical revisit*. In *Proceedings of the 30th International Conference on Neural Information Processing Systems* (pp. 2891–2899). Curran Associates Inc.

Partner Contributions:

Felicia:

- Researched datasets marketing/customer segmentation datasets and proposed possible methods to predict/prevent customer churn.
- Performed initial implementation of classical logistic regression, robust logistic regression, classical K-Means, and robust K-Means on the Telco dataset.
- Explored robust k-means in Georgogiannis's paper, and attempted basic implementation.
- Created and finalized presentation
- Wrote out implementation of project in final report
- Reviewed/edited final report

Jaeyoon:

- Confirmed and finalized appropriate methods for project proposal.
- Performed classical logistic regression, robust logistic regression, classical K-Means, and robust K-Means on the iFood dataset.
- Debugged and fixed issues with robust methods on the iFood and Teleco dataset.
- Reviewed final presentation
- Expanded on introduction, results, and discussion/conclusion in final report
- Reviewed/edited final report