Greyson Corporation Marketing Department

MSBA 605 Case Competition

Renewal Rate Analysis

**Contents:**

1. Executive Summary
2. Case Background
3. Data Preparation and Understanding
4. Modeling
5. Key Performance Indicator Analysis

5.1 Overall Feature Importance

5.2 Response Variable Relationships

5.3 Individual Variable Dependencies and Recommendations

1. Overall Takeaways
2. Proof of Concept – Model Deployment, Lifecycle and Testing
3. Business Value Creation for Greyson Corporation
4. Appendix

**1. Executive Summary**

This research aims to provide Greyson Corporation with a marketing strategy that will boost the customer response rate to the renewal offers for their subscribers. Greyson Corp. has been maintaining a better than industry response rate of over 2% and our research goal is to help them continue to be a top performing business in targeted marketing.

Our objective was to come up with the classification model that can predict the customer response of renewal probabilities, which will eventually increase overall renewals and sales for the company. Our approach was to determine which driving factors effect renewals using historical data provided by the organization.

Our target variable was Renewal which indicates whether the customer renews based on the company's mailing efforts. There were some features with unknown values, which we transformed to enable our model to utilize them in determining the relationship between the predictors and the response. Next the dataset was used on multiple classification models. It was found that the XGBoost model outperformed the other models for the given dataset. XGBoost produced AUC of ROC curve of approximately 90% (0.8895624) accuracy.

We isolated the most significant variables which contributes to the target using varImp function in the XGBoost model. There were seven significant variables ranging from 100 to 30 approximately after executing the algorithm. The top 7 variables of importance were “MonthsSince1stOrder”, “Child6-12”, “Income”, “MonthsSinceExpire”, “DollarsPerIssue”,” ResidenceLength” and “Child13-18” and had the greatest effect on renewals.

We developed a marketing strategy for the organization based on the most critical aspects, which was to emphasize the need of perks such as discounts/rebates, loyalty program etc. to encourage customers to renew their subscriptions. To increase renewals an emphasis should be put on consumers with higher income who also have children in the age group of 6 to 12.

**2. Case Background**

Greyson Corporation is a media and marketing firm that publishes magazine and books, as well as having a broadcasting presence. Their main strength is home and family publications, but the organization has grown to offer a variety of services, which includes, marketing research, communications planning, website advertising, etc.

Greyson Corp. uses data to develop campaigns for new customer acquisition, customer reactivation, and the identification of cross-selling opportunities for products. They send mail to a subscribing customer the version with the print ads identified by their data as being of most interest to that customer.

The problem Grayson Corp faces is the customer response rate to renewal requests sent to magazine subscribers. The organization wants to preserve their competitive advantage of being one of the top achievers in their targeted marketing, even if their response rate is higher than the industry average of about 2%.

Keeping this in mind, we want to improve Greyson's marketing strategy so that they can maintain their position as one of the top targeted marketing performers. We want to accomplish this by utilizing their existing data to develop models that will assist them in predicting which consumers are most likely to reply to their mailing and renew their subscription.

**3. Data Preparation and Understanding**

The dataset provided by Greyson contains 36,000 observations and 38 total variables. Of these variables, 8 are categorical variables and 30 are numeric. The variables given in the dataset represent information that includes customer demographics, homeownership and household details, cancellations, complaints, gift donations, number of magazines paid and how they were paid, expiration and months since last expiration, time from first and last order, payment types and time since last payment, and dollars spent per issue. Seven of the 38 variables have missing values and will have to be assessed prior to deploying the model. There were several nuances within the dataset that were observed while preprocessing the data:

* There are 569 Homeowner observations missing
* There are 5 Home Value observations missing
* There are 1334 Dollars per Issue observations missing
* There are 26 Total Paid Order observations missing
* There are 319 Months Since Last Payment observations missing
* There are 134 Months Since First Order observations missing
* There are 45 Months Since Expire observations missing

Further discrepancies were observed in variables such as:

* Income
* Paid Direct Mail Orders
* Total Amount Paid
* Total Paid Orders
* Unpaid Magazines
* Paid Reinstated Magazines
* Active Subscriptions
* Expired Subscriptions
* Paid Complaints
* Requested Cancellations
* No Pay Cancellations
* Number of Gift Donations

These above-mentioned variables are highly skewed to the right. Their distributions have heavier tails than a normal distribution. However, the data to the right of the median shows a conspicuous variability as well as outliers. Their Kurtosis is greater than 3, therefore, these leptokurtic distributions need to be analyzed carefully due to the higher concentration of extreme values towards the right side of the curve.

Of the before mentioned variables

* Paid Complaints
* Requested Cancellations
* Number of Gift Donations

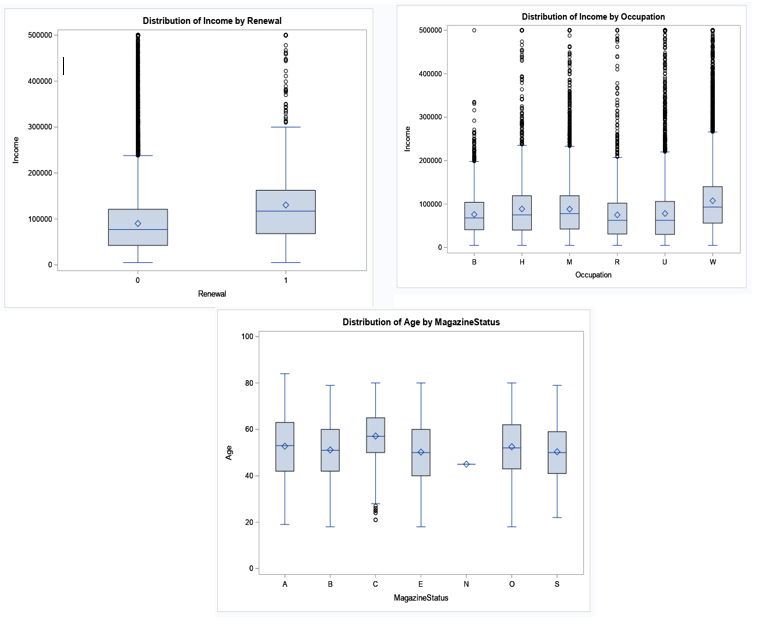
have the highest skewness and Kurtosis values.

On the other hand,

* Age
* Residence Length
* Child 0 – 18
* Years Since Last Order
* Home Value
* Months Since Last Order
* Months Since Expire

reveal distributions with lighter tails than a normal distribution. These distributions seem to be approximately symmetric. However, generally their data to the right of the median seems to be more consistent than the data to the left.  A complete list of these variables with their corresponding statistics can be found in the Appendix.

When the Renewal variable was analyzed alongside Income, the box plot is rather peculiar, due to an interquartile range difference. Moreover, the number of outliers in the distribution of Income by Renewal is concerning. It leads to the conclusion that customers who choose not to renew have a drastic variation in income. Another distribution that is worthy of underscoring is Income by Occupation. Its medians as well as dispersion are reasonably similar across each occupation type. However, the distribution shows a significant number of outliers meaning that income levels also have a high variation by occupation. Finally, the distribution of Age by Magazine Status reveals that the medians are relatively consistent meaning that most of Greyson’s subscribers are between the ages of 40 and 65.



**4. Modeling**

Now that we have explored the data and engineered our features, the next step is to develop an appropriate model for our analysis. Radom Forest and Gradient Boosted Trees (XGBoost implementation) models were fit to determine which produced the highest prediction accuracy. Since our problem is a binary classification problem, the objective was to minimize the log loss by determining the ROC-AUC score. Each model was performed using default grid parameters and utilized five-fold cross-validation to determine an initial accuracy baseline. The models were then set to best tune for optimization. The model that performed the best was the XGBoost gradient boosting tree classifier with hyperparameters: max\_depth = 3, learning\_rate = .025, nrounds = 200, colsample\_bytree = 1, subsample = 1, min\_child\_weight = 1.

A gradient boosted tree is an unsupervised learning technique, which produces a prediction model in the form of an ensemble of weaker prediction models like a decision tree. This technique learns slow by fitting a small tree to the response variable and multiplying it by the learning rate. The residuals of the actual observations minus the predicted observations are then fit to another tree and multiplied by the learning rate again. The results of the two trees are summed and the process continues to iterate in this fashion until the model fits the data well while minimizing overfitting.  The XGboost model was chosen because of the high accuracy of prediction as well as the precautions XGboost takes when it comes to overfitting the training data set. It is also robust to outliers and collinearity, both of which were present in the dataset.

The model resulted in an AUC value of 0.8896 which equates to an accuracy level of 88.96%. Meaning that our model can accurately predict whether a customer will renew their subscription or not 88.96% of the time.

**5. Key Performance Indicator Analysis**

**5.1 Overall Feature Importance**

To determine which of our features had the greatest effect on our target variable, Renewal, a variables of importance plot was produced from the model. The top 6 most impactful variables are listed in the table below with their corresponding importance values. The complete plot can be found in Appendix 3. Our analysis focused on the top 6 variables listed below.

|  |  |
| --- | --- |
| **Variable** | **Importance Value** |
| MonthsSinceFirstOrder | 100.00 |
| Child6.12 | 91.49 |
| Income | 60.65 |
| DollarsPerIssue | 53.98 |
| MonthsSinceExpire | 38.98 |
| Child13.18 | 13.17 |

**5.2 Response Variable Relationships**

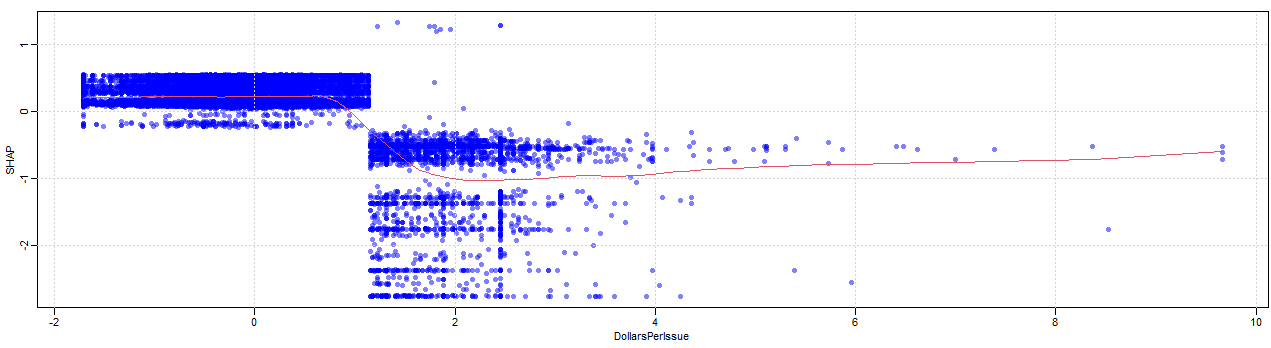
Now that we know which features play an important role in our analysis, we need to understand the relationship between our top 6 features and our response variable. To do this we used a Shapley Additive exPlanations (SHAP) plot. This plot summarizes the general direction and magnitude of impact for the top 6 variables from our model.



The SHAP plot only provides a generalization of the relationships by providing a direction of influence shown. We are interested in determining the actual relationships and whether or not any actionable insights can be derived. To do this, we looked at dependency plots for individual variables.

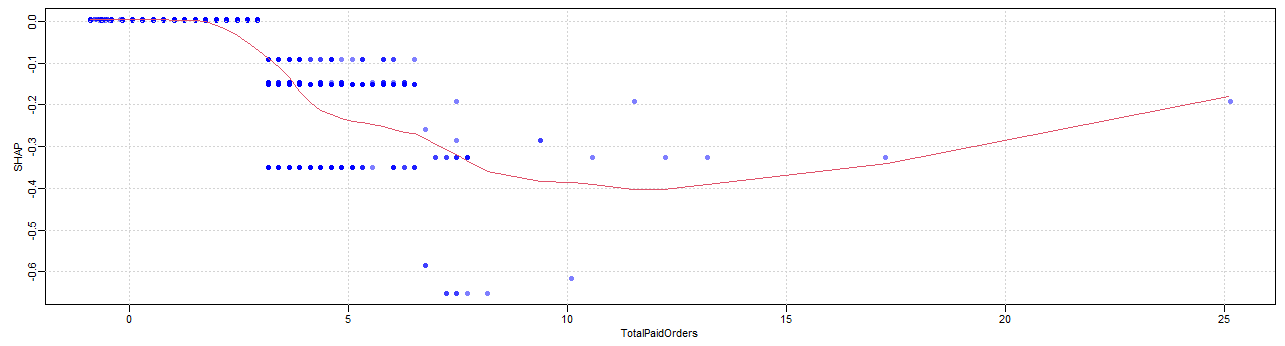
**5.4 Individual Variable Dependencies and Recommendations**

We took a subset of our top 16 most important variables and focused on 6 of the most impactful. For a deeper dive into the evaluation of the top 6 individual variables we used SHAP dependency plots. The plots along with observations and recommendations are illustrated below. The dependency plots for the remaining variables can be found in Appendix 4.



**Observation:** As observed above, a magazine issues that cost more than $1.00 have a negative impact on renewal rates. However, those issues that are valued at less than $1.00 have a minimally positive impact on renewals.

**Recommendation:** It is recommended that discounts per issue be mailed to those customers who have subscriptions containing issues that are either $1.00 or greater than $1.00 per issue. Promotional items can also be advertised towards these customers in hopes of reducing customer churn.



**Observation:** The above plot shows that the greater the number of total paid orders a customer has, the greater the likelihood of nonrenewal. Specifically, once the total number of paid orders is three or more.

**Recommendation:** It is recommended to offer bundled discounts to customers that will have a third paid order or more than three paid orders. A discount of buy 2 magazines and get the third at a discounted rate could work in Greyson’s favor in this case.

**6. Overall Takeaways**

Since the initial stages of our analysis, certain variables displayed a prominent distribution behavior that impacted our model and ultimately our business strategy. The following section encompasses these insights and their connection with our business problem.

* ‘Dollars per Issue’ is a key variable to pay attention to because of the customers Greyson Corporation currently has subscribing. Anyway, of lowering the cost of those paying $1.00-$5.00 per issue would help the renewal rate because of the negative relationship of this variable with the response variable. Why this variable is important is due to those individuals with lower incomes having a higher likelihood of renewing to those with higher incomes in the data set. The current customers look at the costs of a magazine carefully when deciding whether they are going to renew.
* The variables ‘Child 6-12’ and ‘Child 13-18’ are defined as the likelihood of a child of the identified age range being in the home. The predictors have a negative relationship with renewal, as well. The average likelihood of having a child between the ages of 6 and 12 is 42.24% while the average likelihood of having a child between the ages of 13 and 18 is 49.88%. Therefore, a large portion of the customers that are not likely to renew are not convinced that keeping a magazine subscription is important during their busy lives with their families.
* The number of months since the customer’s order and months since their subscription expired also are important to consider. The longer the customer has an active subscription the less likely they are to renew is identified in the variable ‘Months Since First Order’. This likelihood is inherent to any media that is being paid for because customers are excited to read the magazine initially, then after some time, they become bored with reading the same types of ideas every month. The data identifies that the top 25% of the ‘Months Since First Order’ has a value of 13.25 or greater. Therefore, after about a year since the customer’s first order they will likely not renew. This is an actionable time frame to consider reaching out to these customers to prevent nonrenewal.
* The ‘Months Since Expire Variable’ showed an opposite relationship to renewals than the other important variables. The relationship illustrated that within the first 3- 24 months of expiration, a customer is not likely to renew. However, after about three years this relationship changes and individuals have a higher likelihood of renewing. To keep customers who bought a magazine last year, new and different magazine content or ideas should be offered to better keep them engaged. Then for those who have not had a subscription for three years, it is important to reach out to them again.
* The customer’s age shows a relative symmetric distribution. Its median, mode and mean are quite similar. The bottom 25% of the costumers ranges from 18 to 42 which was identified as a promising target. When the ‘Age’ variable decreases, cancelled subscriptions start to slightly increase. This tendency unfolds an area of improvement and potential opportunities. The strategies’ vision will make a priority to appeal to younger generations not only in term of content, but also accessibility and innovation.
* The variable ‘Income’ resonates with our strategy because of its level of importance shown in our analysis. Having a wide range from $5,000 to $500,000 shows signs of a strong variability. Moreover, after delving into ‘Income’, the number of outliers confirmed its high variability. When a customer’s income was analyzed alongside the response variable, their behavior was quite riveting as well. There was a considerable number of outliers for people that renewed and did not renew. Lastly, ‘Income’ belongs to the top 6 variables that affects the response which shows one more time how imperative it is to understand it. The contribution of medium level Incomes impacted the probability to boost the response rate to renewal offers that it mails to its magazine subscribers.

Our analysis sheds light on different areas that Greyson Corp’s could focus to strengthen the customer response rate to renewal offers. Based on our analysis, and from an income perspective, clients with medium level incomes tend to respond favorably to renewals. Nevertheless, the results are not definitive due the number of outliers, suggesting that this behavior could be driven by other elements as well. Furthermore, reducing the probability of having children from 12-18 years old at home reinforces the same tendency. To understand other possible effects, scientific literature related to the media industry was reviewed. According to Scott and Oster, 2003 from the National Bureau of Economics Research, consumers generally experience time inconsistencies. In other words, when the future arrives, a person feels that the action at hand is costly to undertake and does not engage in what she wanted to engage in the day before. One way to target these variables and reduce time inconsistencies is to provide heavily targeted content linking self-growth and meaningful topics that would educate readers in the lower income spectrum about the long-term family benefits of investing in a Greyson’s subscription. It would impel current customers to act in the present to alter the consumption pattern they adopt in the future.

**7. Proof of Concept – Model Deployment, Lifecycle, and Testing**

* **Model Deployment:** The current model was built in R. A one-click application can be created to ingest new data, train the model, provide predictions, and track the model’s performance. This application can be created by Greyson Corporation and deployed in their data server to integrate directly with the data infrastructure they currently have.
* **Model Lifecycle:** This model was built on a historical dataset give to us by Greyson Corporation. Over time, new data will be administered which can cause underlying relationships to change. Because of this, the model’s performance is susceptible to degradation. To prevent this from occurring, the model’s performance should be monitored and retrained using new data, as necessary.
* **Model Testing:** Model testing was performed using the testing data that was split off from the rest of the data at the beginning of the process. 80% of the data was used to train the model, and the other 20% was used to test the model. The testing of this model should always be performed on data that the model was not built off to avoid overfitting, which is critical to an accurate statistical model.

**8. Business Value Creation for Greyson Corporation**

To achieve the goal of increasing renewals we developed a competitive strategy based on the overall takeaways. It would be in Greyson Corp’s best interest to provide a digital platform service for its customers. This will allow Greyson to build an end-to-end solution to conscientiously serve customers and improve operational efficiency.

**Accessibility:** A digital platform service will allow Greyson Corp to provide a wider access to content. In today’s technologically advanced world, accessibility is a necessity. Having access to magazine content, either printed or in digital form will transform Greyson Corp.'s reach to new customers as well as having the ability to provide instant availability to existing customers.

Accessibility not only applies to Greyson Corp’s customers having access to their digital magazine libraries, but also for Greyson to have instant access to their customers. Expiration notices and renewal offers can be quickly and efficiently sent to their customers either via email or within the app itself.

**Consumer Personalization:** With this strategy Greyson Corp can implement an all-access membership plan. This will lead to a more personalized pricing strategy. Rather than paying high dollar per issue, you can now have access to more than one magazine all at an appropriate monthly or annual rate.

**Advertisement Platform:** Promotional items and content spotlights can easily be presented to customers. Greyson can increase revenue from third party advertisers by utilizing their new digital platform. By partnering with companies that complement Greyson Corp’s magazine portfolio, they can generate kickbacks from those companies while retaining customers interested in those products.

These suggestions offer some insights to the conclusions we have drawn from our model and believe that it is in the company's best interest to consider. Each aspect of the proposed digital platform solution goes beyond value creation. It pursues competitive advantage. Our framework focuses on differentiation, inclusiveness, sustainable growth, and a long-term vision. Our solution also bolsters Greyson Corporation to target its market segment effectively to sustain its market leadership position and increase its market share. The strategies provided for our team translate data into actionable steps that emphasize Greyson’s strengths, meanwhile, expanding its scope of opportunities within the industry.

**9. Appendix**

**Appendix 1: Variable Analysis – Summary, Skewness, Kurtosis**

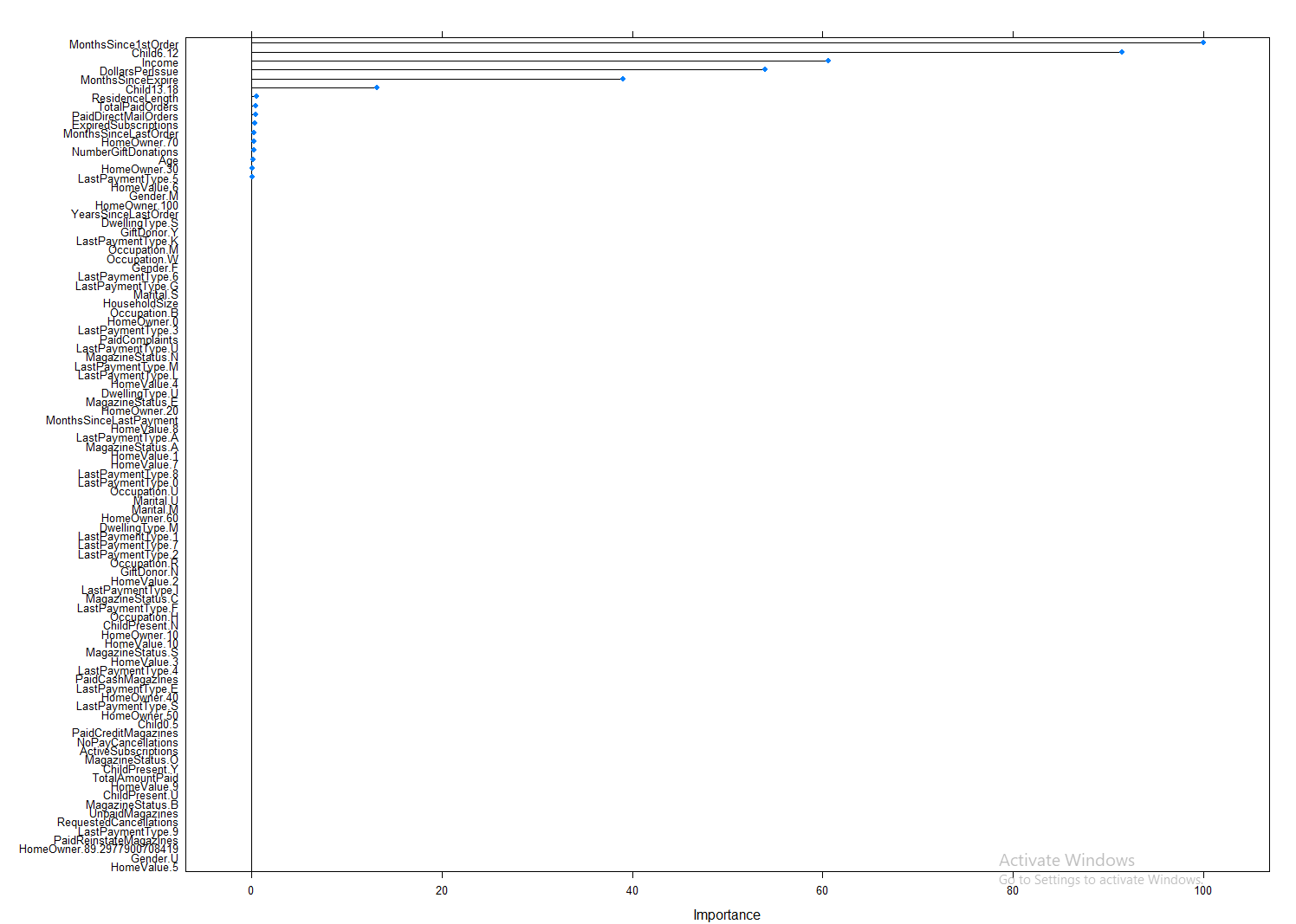
Table

Description automatically generated

**Appendix 2: Feature Definitions**

|  |  |
| --- | --- |
| **Variable** | **Description** |
| CustomerID | Customer identification number |
| Renewal | 1 if customer renewed magazine in response to mailing, 0 otherwise. |
| Age | Customer age (ranges from 18 to 99). |
| HomeOwner | Likelihood of customer owning their own home. |
| ResidenceLength | Number of years customer has lived at current residence. Values: 1 = less than two years 2 = two years 3 = three years 4 = four years 5 = five years 6 = six years 7 = seven years 8 = eight years 9 = nine years 10 = ten years 11 = eleven years 12 = twelve years 13 = thirteen years 14 = fourteen years or more. |
| DwellingType | Identifies the type of residence. S = Single family dwelling unit. M = Multi family dwelling unit, U = unknown |
| Gender | F = female, M = male, U = unknown. |
| Marital | S = Single, M = Married, O = Other (divorced,widowed,etc.), U = unknown |
| HouseholdSize | Indentifies the number of individuals in the household. Arguments are: 1. = 1 person in the household 2 = 2 people in the household 3 = 3 people in the household 4 = 4 people in the household 5 = 5 people in the household 6 = 6 or more people in the household. |
| ChildPresent | Indicates if children are present in the home. Y = child 21 or younger present in the home; N = no child 21 or younger present in the home; U = unknown |
| Child0-5 | Likelihood of child 0-5 years old present in home. |
| Child6-12 | Likelihood of child 6-12 years old present in home. |
| Child13-18 | Likelihood of child 13-18 years old present in home. |
| Income | Estimated income. Ranges from $5,000-$500,000+. |
| Occupation | Broad aggregation of occupations into high level categories. Arguments are: R = retired, W = professional/executive, M = sales/marketing/services/clerical, B = skilled trades / laborers (blue collar type jobs), H = at home (caregivers, unemployed, homemakers), U = unknown |
| HomeValue | The estimated home value in ranges. Arguments are 1-10. 1 = under $50K, 2 = $50K -under $100K, 3 = $100K - under $150K, 4 = $150K - under $200K, 5 = $200K - under $250K, 6 = $250K - under $300K, 7 = $300K - under $350K, 8 = $350K - under $400K, 9 = over $400, 10 = Unknown. |
| MagazineStatus | Identifies the status for a customer based on their magazine business activity. A = active subscriber, B = cancelled subscription due to non-payment, C = cancelled subscription, E = subscription expired within last 3 years, N = gift subscription, O = subscription expired over 3 years ago, S = subscription suspended at request of customer, U = unknown |
| PaidDirectMailOrders | Number of paid direct mail orders across all magazine subscriptions |
| YearsSinceLastOrder | Years since last order across all business lines |
| TotalAmountPaid | Total dollar amount paid for all magazine subscriptions over time. |
| DollarsPerIssue | Paid Amount / Number of Issues Served. Average value per issue (takes the subscription term into account). |
| TotalPaidOrders | Total # of paid orders across all magazine subscriptions. |
| MonthsSinceLastPayment | Recency - # months since most recent payment. |
| LastPaymentType | Indicates how the customer paid on the most recent order. If it was credit order it will contain the billing effort number (how many bills were sent to collect payment). A = cash order, C = mass cancel for non-payment, D = Advanced renewal order cancelled, E = paid via collection agency, F = no billing, G = gift billed, I = online customer payment, K = payment received after order cancelled, L = extra payment, M = step-up financing, S = customer requested order cancellation, U = default paid, 0 = unpaid credit, 1 - 9 = paid credit on ith billing |
| UnpaidMagazines | Number of magazine titles currently in "unpaid" status for a given magazine customer. |
| PaidCashMagazines | Number of magazine titles currently in "paid cash" status for a given magazine customer. |
| PaidReinstateMagazines | Number of magazine titles currently in "paid reinstate" status for a given magazine customer. |
| PaidCreditMagazines | Number of magazine titles currently in "paid credit" status for a given magazine customer. |
| ActiveSubscriptions | Number of different magazines the customer is in "Active" status. |
| ExpiredSubscriptions | Number of different magazines the customer is in "Expire" status. |
| RequestedCancellations | Number of different magazines the customer is in "Cancelled via Customer Request" status. |
| NoPayCancellations | Number of different magazines the customer is in "Cancelled for non-payment" status. |
| PaidComplaints | Number of different magazines the customer is in "Paid Complaint" status. |
| GiftDonor | Yes/No indicator as to whether the customer has given a magazine subscription as a gift. |
| NumberGiftDonations | Number of subscription gift orders for this customer. |
| MonthsSince1stOrder | Recency (in months) of 1st order for this magazine. |
| MonthsSinceLastOrder | Recency (in months) of most recent order for this magazine. |
| MonthsSinceExpire | Recency (in months) since the customer's subscription has expired for this magazine. Negative values represent months until an active subscription expires. |

**Appendix 3: Plot of Important Variables**



**Appendix 4: Dependency Plots**

