Newspaper Churn Analysis with Three Machine Learning Models

I use a subscriber dataset available on the Kaggle platform[1] to predict which subscribers are likely to continue their subscription. The dataset contains 15856 records (n = 15856) of individuals who were all subscribers to an unknown newspaper in California, United States. Some of them quitted the subscription and others remained. The time at which the data is collected is unknown. There are 19 variables (p = 19) that describe the individual. All customers contain a unique SubscriptionID. The dataset contains personal information such as household income, ethnicity, dummy for children, zip code. There are subscription-related data such as weekly fee, delivery period and source channel, which are used to track how the individuals were recruited.

The prediction is important because knowing who is less likely to continue subscription will allow the newspaper to proactively engage them with offers, instead of recruiting new subscribers which is likely to be costly. The exercise is divided into two parts: part A describes the distribution of the main variables of interest, while part B trains three machine learning models: the lasso regression, random forest and boosting.

The R script used for this analysis is available on request.

A. Data Description

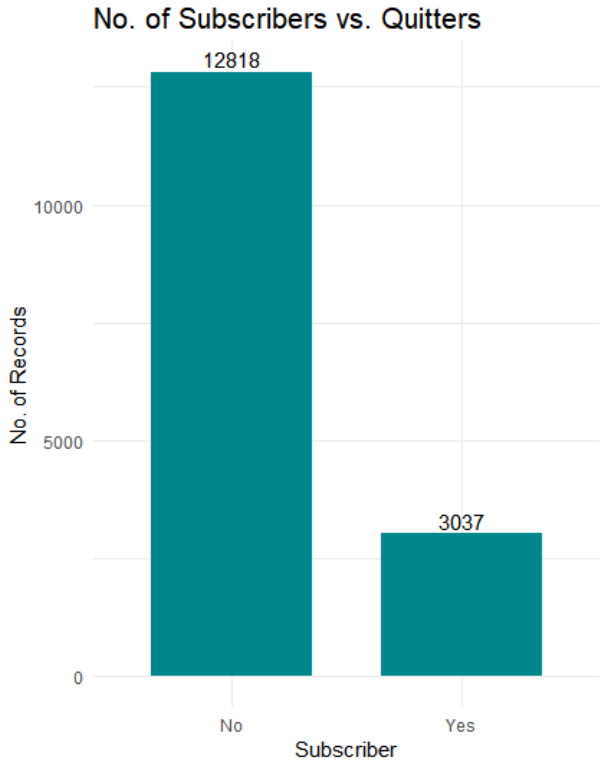

Figure 1 Distribution by subscriber status

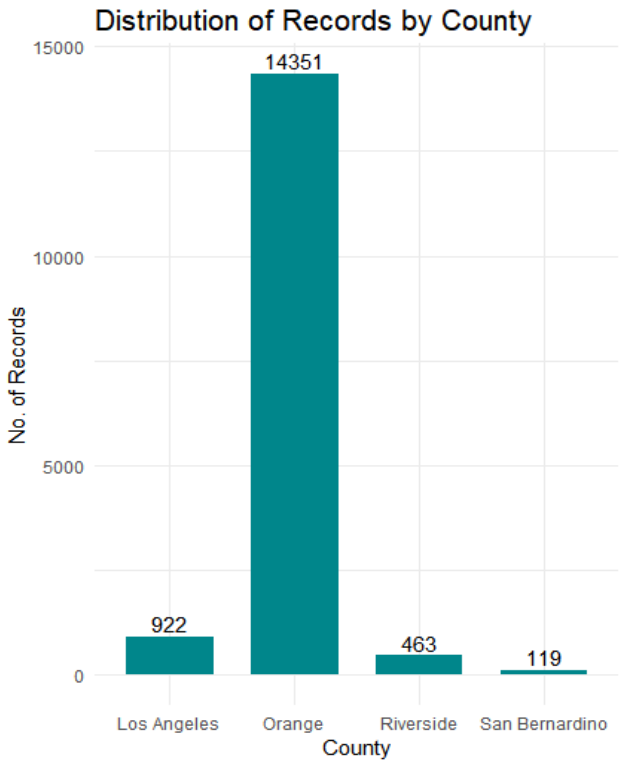

Figure 2 Distribution by county

The dataset is hugely imbalanced, with the no. of quitters (12818) more than 4 times that of current subscribers (3037). 90.5% (14351) of the records are found in Orange County.

[1] Dataset from: https://www.kaggle.com/andieminogue/newspaper-churn/data

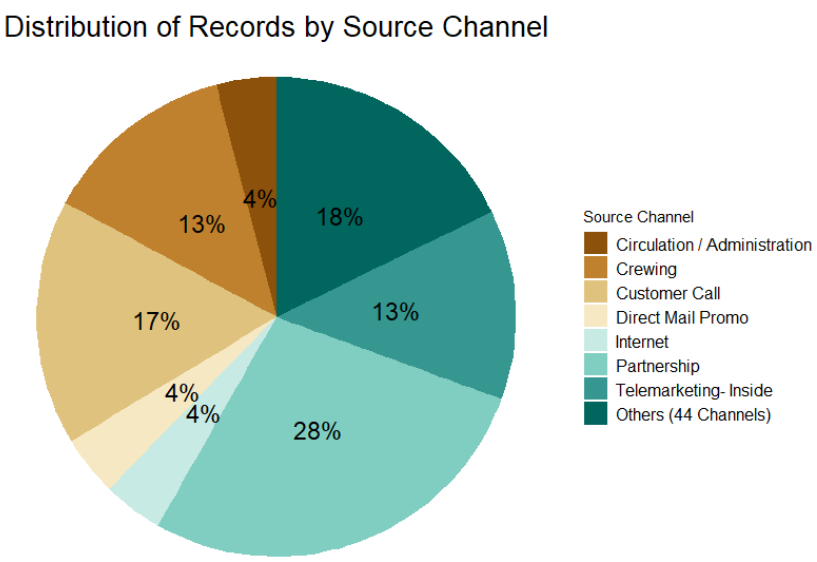

Figure 3 Distribution by Source Channel

There are 51 channels in total through which individuals are contacted. The source channel which was most successful in *recruiting* subscribers is Partnership (28%), followed by Customer Call (17%), Telemarketing – Inside (13%) and Crewing (13%). The data dictionary does not provide much information on what the channels are, but Partnership presumably refers to promoting to individuals who are patrons of a partner institution. Customer call and Telemarketing refer presumably contacting by phone.

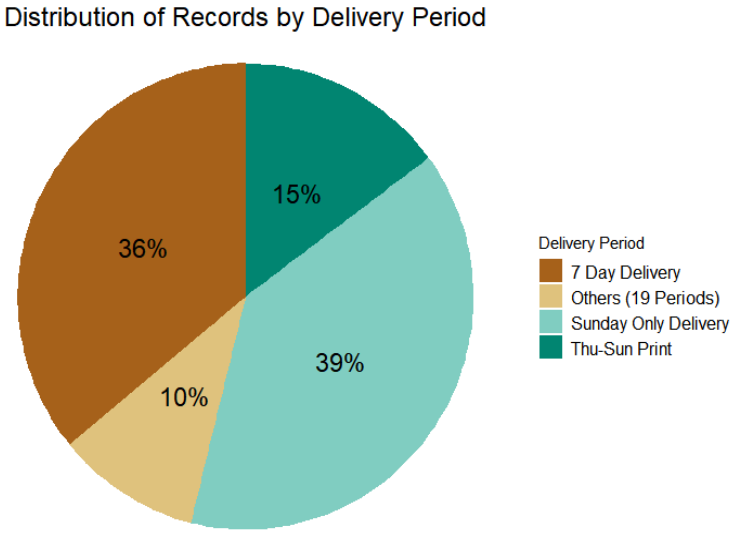

Figure 4 Distribution by Delivery Period

There are 22 delivery periods at which subscribers can choose to receive their newspaper. The top 3 periods dominate 90% of the delivery options. Sunday Only Delivery is the most popular (39%), followed by 7 Day Delivery (36%) and Thu-Sun Print (15%). Online options only make up a mere 5% of the subscription options.

It is worth understanding why the proportion of online readers is so small since online news business has become crucial to the survival of newspapers. Figure 5 shows that price does

not seem to be an issue as more than 1/3 of online readers pay less than $2.99 weekly. Without prices offered by other online news provider at the time when the data is collected, however, we are unable to benchmark this figure against competitors.

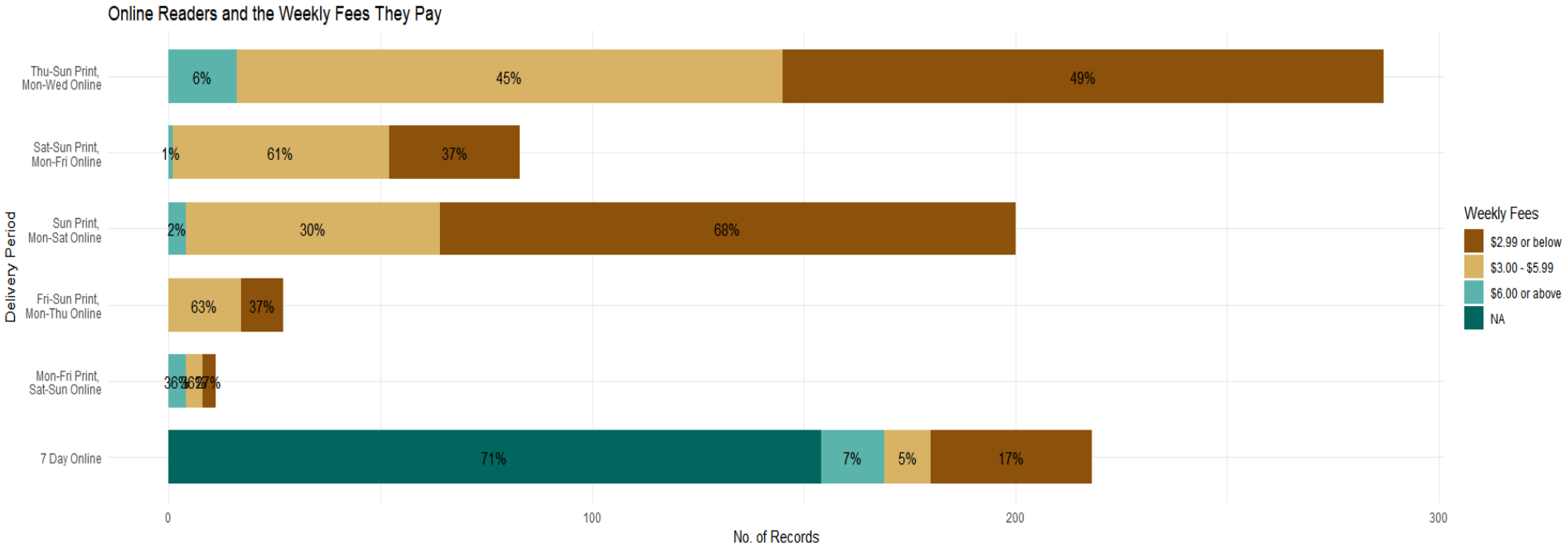

Figure 5 Online readers often pay small weekly fees

Two more factors could possibly be contributing to the small proportion of online readers: 1) The newspaper does not have an appealing user interface for its online version 2) The dataset is dated before the widespread digitalisation of newspaper. If it is the former, the newspaper should seriously consider improving the online user experience for its publishing business to remain relevant.

It is also important to note that online subscription pattern can be correlated to demographic makeup. Older age groups, for example, might prefer printed over online news. As mentioned, the dataset is heavily concentrated in the Orange County. The dataset might therefore be simply reflecting the preference of this population instead of the behavioural pattern of the overall reader base.

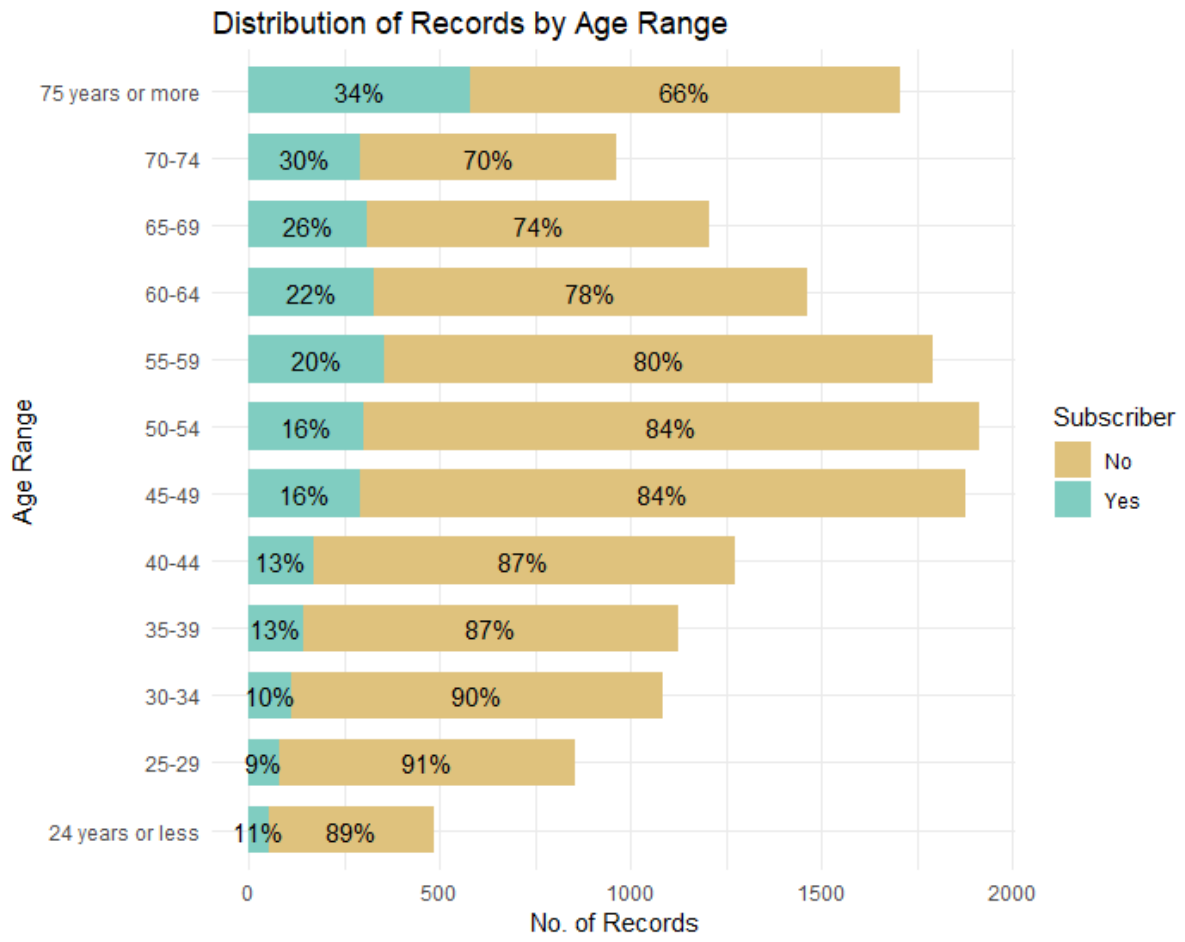

Figure 6 Distribution by Age Range

Figure 6 indeed shows that the typically "internet-savvy population" (age 24-40) only makes up 22.5% of the reader base. Most of the readers the newspaper recruits are within the 40-64 age range, which are less likely to be online readers. Nonetheless, among readers below 40 years old only 5.2% have chosen a partial or full online subscription. This supports the hypothesis that the company has only attracted the non-online readers among the younger generation. More marketing efforts and business development are needed to expand the company's online business.

Since many of the variables are not continuous, neither a simple correlation matrix nor principle component analysis can be constructed directly for data exploratory purposes. A few categorical variables, including ethnicity, weekly fees and household incomes, are therefore taken to build stacked bar charts to see if they display a pattern. The pattern of age range is particularly interesting and is shown above in Figure 6. While quit rate is high among the young readers, it decreases steadily with older age groups. Income displays a similar, but much less marked decrease in churn rate for high earners.

The only continuous variables, years of residence and no. of reward programs both display less than 0.2 correlation with subscription status.

B. Model Predictions

The exploratory analysis above assessed the current performance of the reader recruitment. It also helped me to select a subset of variables for models. Now let's turn to the three ML models to predict the likelihood of an individual remaining.

B.1. Lasso Regression

Among parametric models, I chose the lasso regression as it is more suitable for a regression with potentially a lot of noise. The lasso regression looks for coefficients that minimize the residual sum of squares (RSS) and a penalty term (see formulation below). The lasso penalty term shrinks coefficients of those variables that are not important to 0. When there are many variables (p) relative to the number of observations (n), using the lasso can reduce variance at the cost of only a small increase in bias. Below is the formulation:

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j| \;=\mathrm{RSS}+\lambda\sum_{j=1}^{p}|\beta_j|.$$

Like any other parametric model, lasso regression requires model specification. Based on variables which I saw showing a subscription pattern in the exploratory analysis, I formulate the model as follows:

$$Subscriber\ Status = \beta_0 + \beta_1\ Household\ Income + \beta_2\ Home\ Ownership + \beta_3\ Ethnicity + \beta_4\ Children + \beta_5\ Years\ of\ Residence + \beta_6\ Age\ Range + \beta_7\ Zip\ Code + \beta_8\ Weekly\ Fee + \beta_9\ Nielsen\ Prizm + \beta_{10}\ Source\ Channel + \beta_{11}\ Reward\ Program + \varepsilon$$

where *Subscriber Status* is a binary variable which takes on the value of 1 if the reader remains with the newspaper and 0 if s/he quits. *Household Income* is the annual income of the individual's household ranging from under $20,000 to more than $500,000. *Home Ownership* is a dummy variable indicating whether an individual is a renter (=1) or an owner (=0). *Children* is a dummy variable taking on a value of 1 if the subscriber has children and 0 if otherwise. Ethnicity has been consolidated from 73 categories to 5 categories: White, Native Haiwaiian or Other Pacific Islander, Hispanic, Asian, Black or African American. These 5 categories are commonly used in US Census. *Age Range* is measured in intervals from 24 years or less to 75 years or more. Zip code has been reduced from 6 digits to the first 3 digits to group neighbouring readers together. *Weekly fee* refers to the amount of money a subscriber pays per week and ranges from $0 to $9.99. *Delivery period* is the frequency (how many days a week) and medium (paper or online) at which a subscriber receives their subscription. *Nielsen Prizm* is a market segmenting method which is not explained in the dataset but displays a pattern in subscription when I built a stacked bar chart. *Reward Program* is the number of offers used by the individual.

Since all chosen variables except reward program and years of residence are categorical, dummy variables are created. This gives me a sparse model matrix of 73 variables (p = 73). After removing rows with NA entries, I am left with 14461 observations. I use around 75% (n_{train} = 10845) of the sample to train the model. The remaining 25% (n_{test} = 3616) is used for

testing the accuracy of the model. The 15 variables with the highest (absolute) values of coefficients are as follows:

Variable	Coefficient
Weekly Fee $10.00 - $10.99	2.363
Source Channel: Partner	1.509
Weekly Fee $8.00 - $8.99	0.942
Weekly Fee $1.00 - $1.99	-0.812
Weekly Fee $0.01 - $0.5	0.668
Weekly Fee $3.00 - $3.99	0.571
Weekly Fee $4.00 - $4.99	-0.496
Source Channel: Internet	0.491
Source Channel: Circulation/ Administration	-0.481
Deliver Period: Others	-0.480
Weekly Fee $6.00 - $6.99	0.445
Weekly Fee $2.00 - $2.99	-0.339
Years of Residence	0.259
Source Channel: Direct Mail Promo	0.234
Reward Program	0.199

Table 1. Variables with the highest coefficient values in the lasso model[2]

Positive values of estimates mean that the variable increases the likelihood of a person *remaining* as a subscriber. Different values of weekly fees are reasonable predictors on whether a reader will continue their subscription. From figure 7 we can see that those that are paying the highest amount of weekly fees ($10.00 - $10.99) are least likely to quit. Those that are paying less are more likely to quit. Quit rate is particularly high among the $1.00-$1.99 and $2.00 - $2.99 groups. Though it should be noted that a weekly fee of $3.00-$3.99 predicts a higher likelihood of continual subscription.

[2] No significance test is available as the null hypothesis is no longer that these variables has no significant effect on the dependent variable (more technical details please check http://web.stanford.edu/~hastie/StatLearnSparsity/)

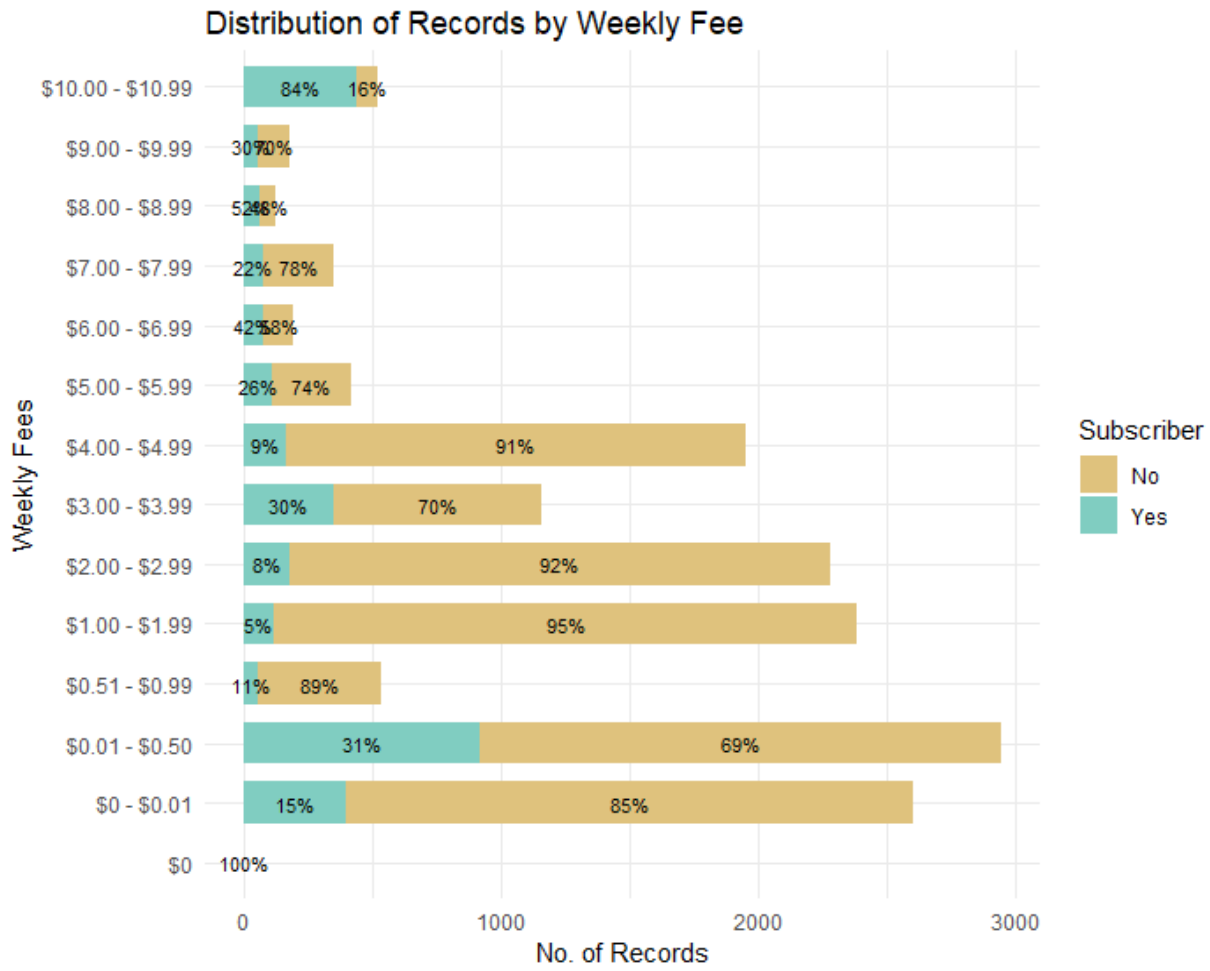

Figure 7 Stacked bar chart to show how lasso coefficients correspond with data pattern

The coefficients on four variables – *Household Income, Home Ownership, Children* and *Zip Code* are reduced to 0 in the lasso regression. This is rather unexpected because one would expect higher income earners to remain subscribers. As I expect zip code to correlate with income, the 0 effect of zip code suggests that income might indeed predict poorly on whether a reader would stay.

The **training error** (using validation set approach) is **16.4%** and the **test error** is **17.0%**, showing that the model did not overfit the data and predicts the subscription outcome rather reliably. Since lasso is the most interpretable among the three models, we will refer to it for recommended actions.

B.2 Random forest

Random forest is an extension of the decision tree method. Decision trees put observations into different classes according to the most commonly occurring class of the training observations in the terminal node to which it belongs. Using the graph below as an example, a terminal node is a circle that is found at the end of the tree. For instance, individual observations are sorted after several splits into the node circled in red as they share same characteristics (i.e. transaction < 18 and small scale). There could be some actual customers who belong to the churn group but get sorted into the red node as well. But since there are

more actual no-churn customers than churn customers on average in that node, the node is labelled “no-churn”.

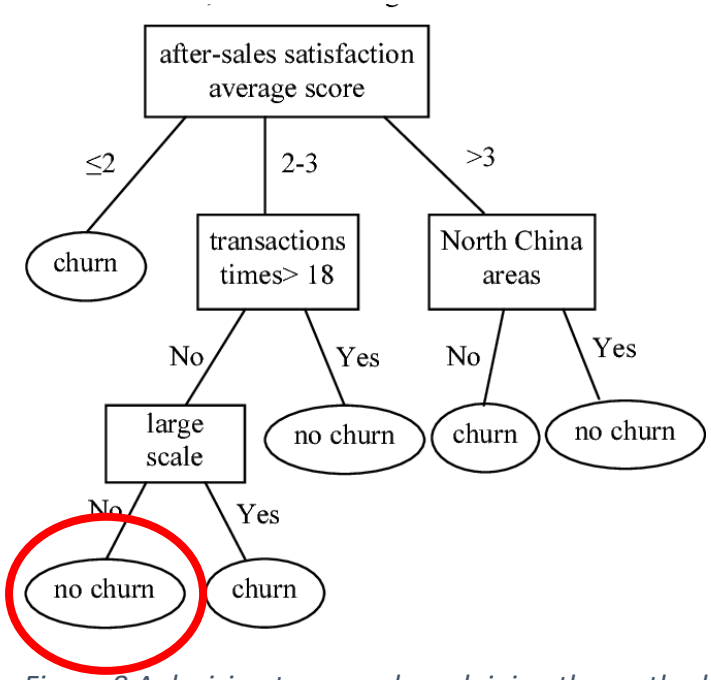

Figure 8 A decision tree graph explaining the methodology[3]

In a random forest, decision trees are built with repeated sampling from the original training data (a process referred to as bootstrapping), and with one predictor chosen from a random sample of *m* predictors at each split. The process is known as bagging if m = p, i.e. when the predictor is chosen from the whole set of predictors at each split. The results are then averaged to see which class has an individual been most frequently assign to. In that case trees are decorrelated and variance can be reduced.

[3] Source: Ma, H., & Qin, M. (2009). Research Method of Customer Churn Crisis Based on Decision Tree. *2009 International Conference on Management and Service Science*, 1-4.

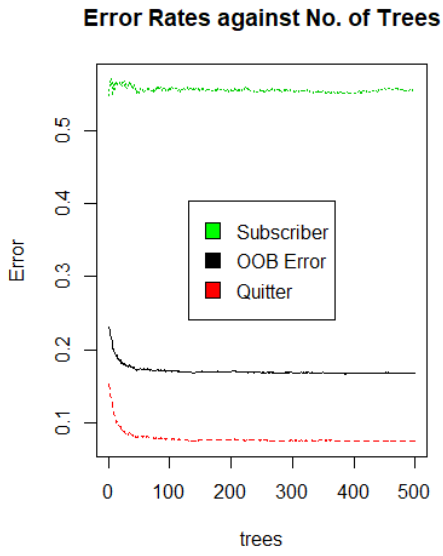

Figure 9 No. of Trees is chosen based on the OOB Error as it levels off at around 21 trees

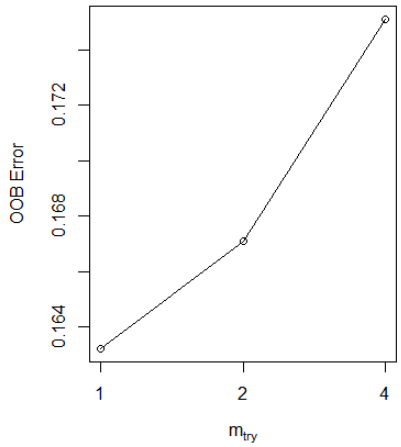

Figure 10 m is the number of predictors sampled at each split. m = 2 gives a low OOB Error while not overfitting

I continue with the variables used in the lasso regression but removed *Household Income, Home Ownership, Children* and *Zip Code* to reduce noise. Since no dummy needs to be created (random forest in R can handle categorical data), I have only 8 variables in total. After using 10-fold cross validation to choose hyperparameters, I chose 21 trees (see Figure 9) with 2 predictors (see Figure 10) taken at each split. The size of training data is the same as the lasso regression ($n_{training}$ = 10845).

Unlike regressions, random forest does not have a coefficient indicating whether a variable will increase or decrease the likelihood of continual subscription. However, the variable importance table tells us which variables have been most useful in decreasing the Gini index. The lower the Gini index, the higher the node purity.

Variables	Mean Decrease Gini
Weekly Fee	666.97
Years of Residence	415.61
Age Range	306.82
Nielsen Prizm	246.14
Source Channel	235.01
Reward Program	225.04
Delivery Period	131.00
Ethnicity	103.75

Table 2. A variance importance table of the random forest model.

The values tell a similar story as the lasso regression. *Weekly Fee* is still an important predictor of subscription outcome. *Years of Residence* is not an important predictor in the lasso regression but comes up second on this table. The importance of Age Range as a predictor strengthens the assumption from figure 6 that age range does display a subscription pattern.

The **training error** (out-of-bag error) is **16.6%** while the **test error rate** is **15.5%**, a slight improvement from the OLS model.

B.3. Boosting

Boosting is another tree-based method. Rather than taking repeated samples from the original training data, trees in a boosting approach is grown sequentially. A tree is fitted using the residuals from the previous fit rather than the outcome variable Y. The process learns slowly and is less likely to overfit the data.

The size of training and test sets is identical to the previous two models. I used the same 8 variables as the random forest model. Using the validation set approach to randomly search for hyperparameters, I fit a boosting model with ntrees = 900, interaction depth = 5, shrinkage = 0.1, minimum number of observations in node = 12, and bag fraction = 0.8. The smallest training error obtained is only 7%, but I reckon this model might overfit the data and therefore opt for a model with a higher misclassification error and the hyperparameters above.

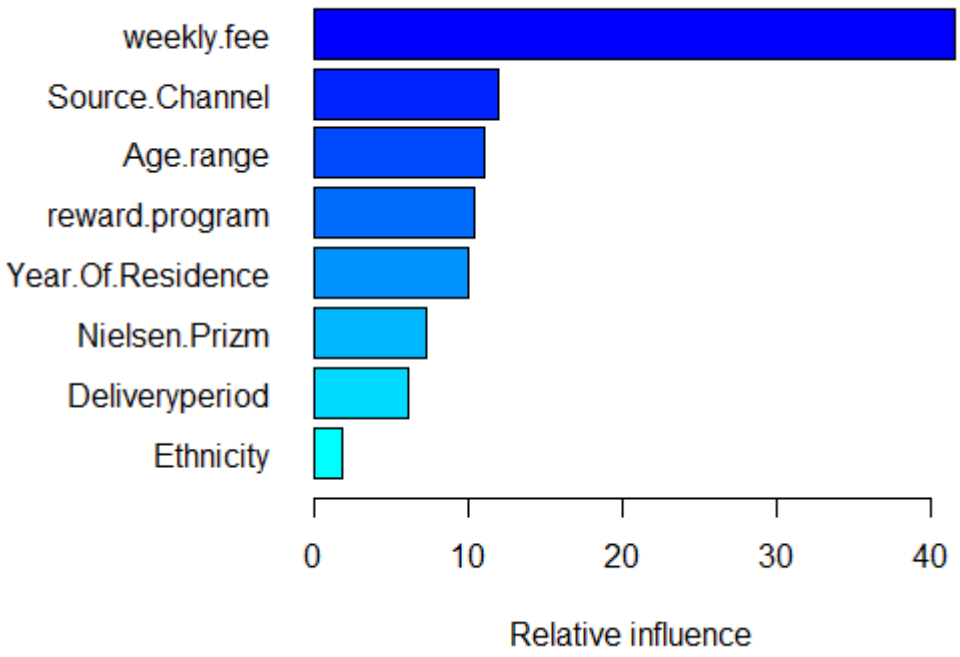

Figure 11 Variable Importance in the boosting method plotted

Weekly Fee continues to be the most important predictor. *Source Channel*, which is shown to be important in lasso but not so much in random forest, comes second. Like in random forest, *Age Range* continues to exert influence on the prediction.

The training error (OOB error) rate is **10.1%,** whereas the **test error** rate is **15.4%**. A slight overfit but still performing well.

C. Comparison and Recommendations

The training and test errors of the three models are as follow:

	Lasso	Random Forest	Boosting
Training error	0.164	0.166	0.101
Test error	0.170	0.155	0.154
Overfit?	Slightly	No	Slightly

Table 3. Training and test errors for the three models

The performance of the three models can be compared in their Receiver Operator Characteristic (ROC) curves as follow:

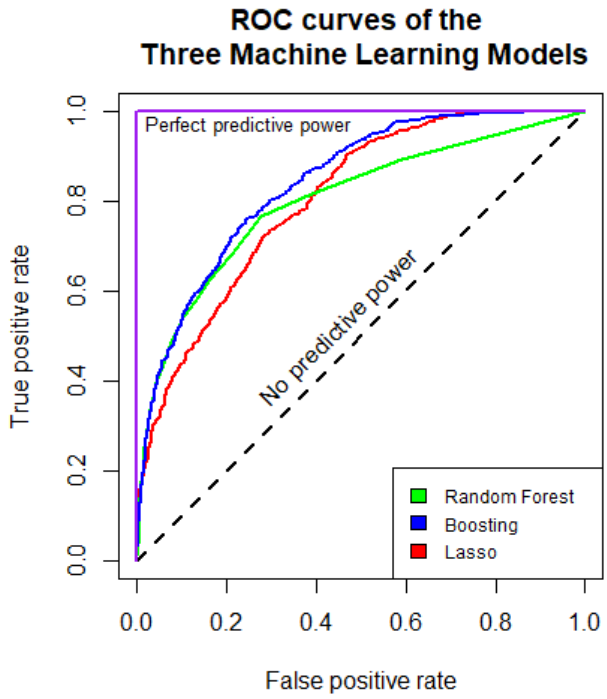

Figure 11 ROC curves showing precision and sensitivity of each model

ROC curve shows the trade-off between true positive rate and false positive rate. True positive rate and false positive rate in this context are defined as follow:

$$True\ Positive\ Rate\ (TPR) = \frac{Actual\ Subscribers\ who\ are\ predicted\ to\ be\ Subscribers}{Predicted\ Subscribers}$$

$$False\ Positive\ Rate\ (FPR) = \frac{Predicted\ Subscribers\ who\ are\ actually\ quitters}{Actual\ Quitters}$$

TPR and 1-FPR are also known as the model's sensitivity and specificity correspondingly. The lines indicate that the more successfully our prediction captures actual subscribers in our prediction, we are also more likely to wrongly predict some quitters as subscribers.

A perfect model will have a line that traces the y-axis to the top of the chart then turns right to trace off the top part of the chart (as coloured in purple). A model that falls below the diagonal line has no predictive power.

Boosting has the highest predictive power, with random forest and lasso each qualifies as a runner-up depending what measure we use. If we use test error as an accuracy, random forest outruns lasso slightly by 1.5%. If we are more interested in predicting who will quit and hence take actions, lasso might be a better model since it has a lower increase in false positive rate given every 1% increase in true positive rate.

Model	Replications	Time Elapsed (sec)
Boosting	10	67.75
Lasso	10	38.11
Random Forest	10	3.74

Table 4. Speed of each training execution

From a pure execution perspective (ignoring the time needed to tune the hyperparameters), random forest is the fastest model with only 3.74 seconds needed, on average over 10 iterations, to train the model with optimised hyperparameters. Lasso requires 38.11 seconds on average. Boosting understandably takes the longest (67.75 seconds) as it learns slowly by fitting trees on the residuals of a previously built tree.

Here are three types of recommendations from this exercise:

a. Model for churn prediction
   Boosting is the best method among the three by a small margin (test error = 0.154). However, if the dataset is large (e.g. >200,000 instead of just 10,000), the analyst can also consider using random forest, since it runs faster with negligible increase in test error. And if having an interpretable model (i.e. one can point out the direction of each variable's effect on subscription outcome) and predicting quitters is the most important, lasso is the preferred model.

b. Business-specific recommendations
   There are three recommended measures:
   1. Focus on reader who are paying less
      It is clear from the three models that weekly fee is the most important predictor of continual subscriber status. In particular, the lasso regression suggests that those which pay the most fees ($10.00-$10.99) are most likely to stay and vice versa. The less-paying readers might have left their subscription because the trial or promotion period has ended.

      Therefore, if through a cost benefit analysis, the business decides it is more profitable to retain current readers than to reach out to new readers, it should actively engage reader groups which are paying less with more discounts and promotion.

      It is worth noting that income indicators such as household income and zip code are poor predictors of subscription outcome. The observation suggests that this newspaper is a price elastic rather than an income elastic good.

2. Expand on recruitment through partnership, internet and direct mail promo.
   The three models have also shown that source channels through which readers were recruited in the first place predict reasonably well whether they will continue their subscription. Interestingly, the lasso regression shows that internet readership recruitment has a positive effect on the subscription outcome. Although the mechanisms are unknown as I do not have details about how the source channels work, the three models suggest the company should expand its recruitment through partnership (which is already where it recruits most of its readers from), internet and direct mail promotion.

3. Attract long-time residents.
   Years of residence is also an important predictor in all the three analyses. The positive sign on its lasso coefficient suggests that readers which have been a long-time occupant of their residence are more likely to stay as subscriber. However, years of residence could be correlated with age which might again just reaffirm that the news is better at attracting older readers. Still, the newspaper is recommended to find channels where they can attract long-time residents, such as through promotions at townhalls or other community events.

4. Expand online news business
   Figure 6 shows that the newspaper has a small population of readers below 40 and is very popular among readers aged 75 years or more. Since readers under 40 are typically assumed to be online news readers but only 5% of the under 40s in this analysis have opted for online news, I believe the newspaper is indeed capturing the "non-online readers" among the younger generation. Still, the paper's political stance is a potential factor that cannot be ruled out. If the weak online presence is the case, the company should focus on expanding its online business to attract more high-spending and internet-savvy millennial readers.

   At the same time, a major reason why the newspaper loses more short-time occupant might be due to its print nature. This problem can be solved by moving its business online.

c. Data Collection
   To improve prediction precision, it is recommended to record exact numbers wherever possible. Numerical data like weekly fees are best recorded in exact numbers instead of a range to allow for more exploratory data exercises such as principle component analysis that could perhaps yield more consumer insights. On the other hand, categorical data such as ethnicity or city is best recorded in broad categories to reduce computational demand and variance in model predictions. If included in a flexible model, variables like ethnicity which only have only a few observations in each category can lead to overfitting, since the predicted values will be highly sensitive to any small change in data.

D. Limitation

Given that this is just an anonymous dataset from Kaggle, there is a lot of limitations with this analysis. First, the dataset is biased in terms of location and will not reflect the *company's performance* on a whole if its business area extends beyond the Orange County, where 90.5% of the records are found. Second, the *consumer behaviour* reflected might only be peculiar to the Orange County and should not be generalised to the business's operations in other areas until more data is obtained. I also have very little information about the company's business operation that will allow me to give more specific recommendations – for instance, details about the fee structure (e.g. why are people with the same delivery option paying different prices? Are some of them offered trials?) and source channels (e.g. What does partnership mean? Did they advertise online if people are recruited through "internet"?) are unknown. It will also be helpful if I can understand the correlation between different variables through a principal component analysis or clustering, were more continuous variables recorded.

E. Conclusion

I fitted three machine learning models to predict who is likely to continue subscription from a newspaper reader base in California. With around 10,000 training data, my best model predicts subscription outcome with around 85% accuracy. I found that weekly fee is the most important predictor of subscription outcome and that the newspaper should focus on low-paying readers if the goal is to retain current customers. More importantly, I found that the newspaper has little appeal to younger readers and has a weak online subscription base, which implies a weak online business. The latter not only reaffirms the former but also results in a loss of readership from frequently moving professionals/ short-term tenants to whom paper delivery is inconvenient.