STAT642 Data Mining for Business Analytics

CUSTOMER CHURN ANALYSIS

Group 3 Camila Pareja, Abhinav Sapru,
Jiameizi (Violet) Yao, Yue(Alex) Fu

Drexel University

TABLE OF CONTENTS

Introduction

Customer churn is an expensive problem that all businesses face. We used the telco churn dataset from Kaggle to predict customer churn, identify features that play a key role in predicting process, and present some recommendations based on our findings.

We identified the Decision Tree and Support Vector as two models that could help us accurately predict customer churn.

Data

We have a data about 7043 customers. There are 21 variables in the dataset - one ID field, 3 numerical independent variables, 16 categorical independent variables, and a two-level dependent variable, which identifies whether a particular customer has churned or not. **Exhibit 1** lists all the variables and a brief description of each.

Exploratory Analysis

We explored the structure and summary of all the variables in the dataset.

Dependent Variable

The dependent variable is Churn. It has two levels. As shown, in the following distribution, the two classes are imbalanced.

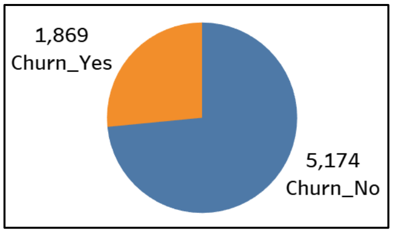


Figure 1: Distribution of Churn in the dataset

Independent Variables

Before we start building our models, we first looked at the distribution of the depedent variables. Some of them are presented below.

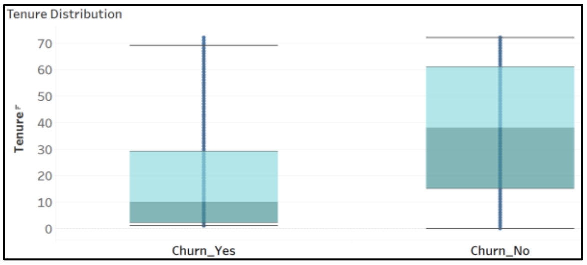


Figure 2: Distribution of Tenure

Tenure in our dataset refers to the time length (measured in months) customers stay with the company. As shown from the box plot above, customers who churn, have lower tenure – they have been customers for a shorter duration.

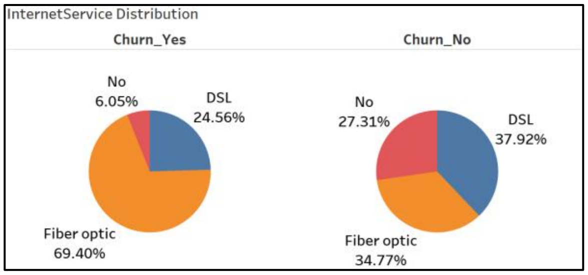


Figure 3: Distribution of Internet Service

There are three classes of internet services, fiber optic, DSL, and no internet service. Based on the pie chart above, the majority of customers who churn use fiber optic as their internet service.

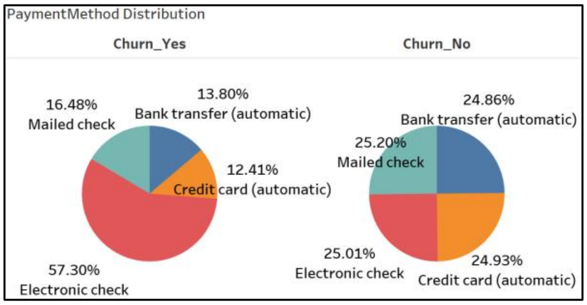


Figure 4: Distribution of Payment Method

Customers can choose four different ways to make a payment. The distribution chart from above, shows customers who churn used to pay with electronic check more.

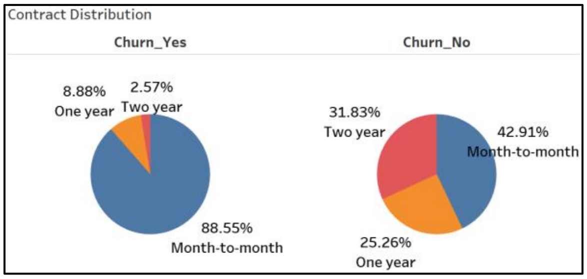


Figure 5: Distribution of Contract Type

As for the length of contracts, the majority of customers who churn have month-to-month contracts. Customers with two-year contracts churn at a lower rate.

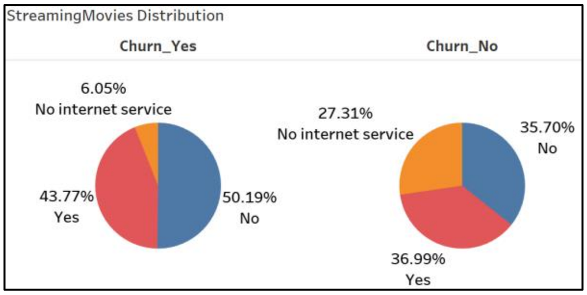


Figure 6: Distribution of Steaming Movies

We included streaming movies as one of the selected variables in our decision tree model. There are slightly more customers who do not subscribe to the streaming movies service and churn.

Data Quality

Missing values

The original data has 11 missing values for the *TotalCharges* variable. Since the corresponding rows are a small portion of the whole dataset, we deleted these rows from our data.

Outliers

None of the three numeric variables have outliers.

Pre-processing

ID variable

The data has an ID field which is unique for each row of data. We deleted this variable.

Categorical Variables

We created dummy variables for all the levels of categorical variables.

Resampling

Despite the class imbalance, we decided not to balance the data as the methods we selected either do not require it (decision tree) or have parameters that can be used to ensure that the imbalance does not cause the model to be biased towards the majority class (support vector machine).

Feature Selection

Correlation Analysis

We checked the correlation between the dependent variables to identify redundant variables.

There were 9 correlation pairs where the magnitude of the correlation coefficient was greater than 0.7. We deleted one variable from each of these pairs. **Exhibit 2** lists the removed variables.

Random Forest

We still had 21 dependent variables. The *MeanDecreaseGini* plot for random forest helped us to identify the features which contributed the most to predicting churn.

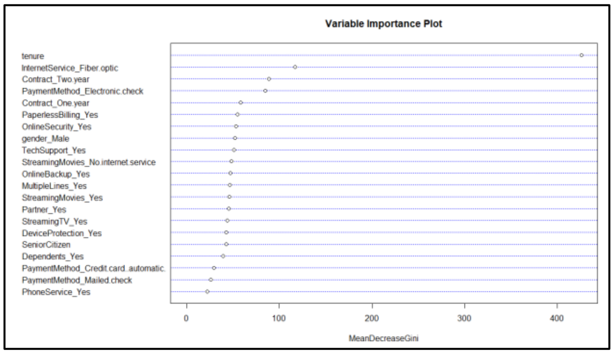


Figure 7: Random Forest Variable Importance Plot

We built the random forest model using the training data. The result shows that *tenure, InternetService_Fiber.optic, Contract_Two_.year, PaymentMethod_Electronic.check,* are the top four variables that can help predict customer churn.

Analysis

Model Evaluation

Train/Test

We randomly split our data with 80% of the rows in the training set and 20% in the testing set.

Metrics

Since our data is imbalanced, accuracy is not a good measure of model performance. As we want to maximize the number of churn cases identified, we decided to use Recall and F1 scores as the best metrics for evaluating our models.

Decision Tree

Since we are looking to predict the reasons for customer churn, in order to provide valuable recommendations to help companies improve customer retention, decision trees is an excellent method to classify these outcomes within a simple structure of branching decisions. These branching decisions allow us to easily interpret the causes of customer attrition, allowing us to develop recommendations by carefully analyzing variable importance and the results within these branches.

Tuning & Variable Selection

After testing several procedures such as Grid Search and basic DT classification with no hyperparameter tuning, we received better results by using Random Search with hyperparameter tuning. We created an alternate model with the most important variables using the results in **Figure 7** and **Figure 8**.

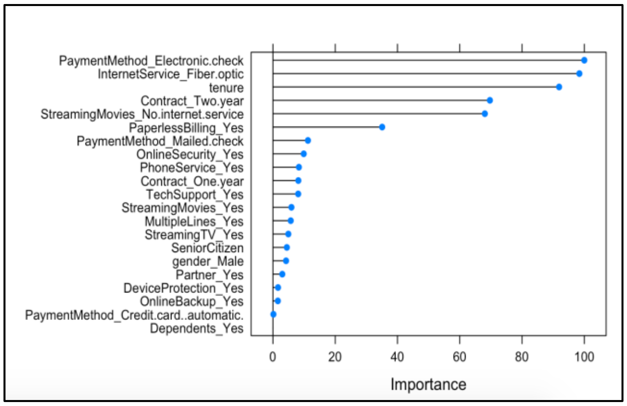


Figure 8: Variable Importance Plot

Model Testing

For our first model, we included all 21 variables, in order to build a model following our plot suggestion in ***Fig 8***. Since we did not receive satisfactory results for Kappa, Recall, and F1 scores. This model suggested that our data may be overfitting.

Therefore, we decided to build a second model only with our top 4 variables (***Table 1)***. Model 2 presented much better results overall, with scores in Kappa, Recall, and F1 close to each other between our training and testing data, displaying a good model fit. Hence, we concluded that model 2 best represented our data and provided us with the most relevant variables that will allow us to formulate effective recommendations to avoid customer churn.

Relevant Variables |Model 2
Tenure
Payment Method_Electronic.check
InternetService_Fiber.optic
Contract_Two.year

Table 1: Top 4 Variables used in Model 2

Model Performance

The performance comparison between Model 1 and Model 2 is shown in ***Table 2***. We can clearly see that our second model performs much better, with only our top 4 most relevant variables in it.

Metric	MODEL 1		MODEL 2	
	21 Variable		4 Variables	
	Train	Test	Train	Test
Precision	0.74	0.65	0.67	0.65
Kappa	0.55	0.40	0.44	0.41
Recall	0.59	0.46	0.49	0.46
F1	0.66	0.54	0.57	0.54

Table 2: Model 1 vs Model 2 results

Decision Tree Interpretation

After selecting our best model (Model 2), we used the *rpart.plot package* to exhibit a tree plot with only our most relevant and informative variables show in **Exhibit 3**. This tree plot provided us with the following information:

1. ***Tenure*** seems to be the most relevant variable and an important indicator for customer churning.
2. ***Electronic check payment methods*** seem to be inefficient and subject to customer churning.
3. ***Streaming Movies*** is the least mentioned variable in our tree plot. However, it provides us with information that at least half of customers who churn do not stream for movies and that having no internet service is quite irrelevant for customer churn.
4. The combination of Tenure of less than 10 months, and the electronic check payment method seems to increase the chances of customer churn.

Support Vector Machine

We selected SVM as our second model because it is considered one of the best "out of the box" classifiers. Additionally, it can accommodate non-linear decision boundaries. It works by identifying an optimally separating hyperpplane between the two classes.

Class Weights

Since the data is imbalanced (ratio of not-churned to churned is 2.7:1), we assigned a weight of 2.7 to the churn class to balance the influence of both the groups on the model.

Kernel Selection

We compared the performance of a linear and radial kernel on the test data. The performance of both kernels was similar on the training and test data in terms of F1 score and Recall. We decided to proceed with the radial kernel as it allows for a more flexible decision boundary.

Models

We built two models. One with the 21 variables left after we removed the highly correlated variables. The other that just uses the top 4 variables identified by our random forest variable importance analysis.

The F1 score was identical for both these models, and the 4-variable model has a better value of recall. The performance of both these models is presented in the following table.

Metric	MODEL 1		MODEL 2	
	21 Variable		4 Variables	
	Train	Test	Train	Test
Precision	0.58	0.53	0.48	0.48
Kappa	0.55	0.46	0.41	0.41
Recall	0.86	0.78	0.83	0.82
F1	0.69	0.63	0.61	0.61

Table 3: Results of Two models

Principal Component Analysis

For the 4-variable model, we transformed the data into its principal components. The first two principal components account for 75% of the variation in the four variables.

Figure 9 shows the plot of the original variables versus the principal components.

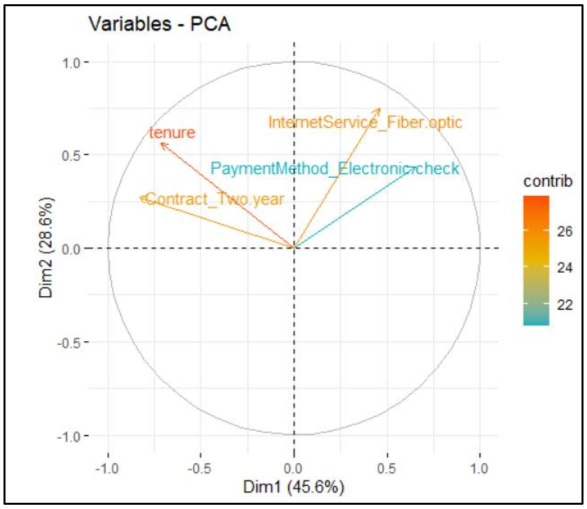


Figure 9: PCA plot for top variables

This figure shows that

- High values of PC1 corresponds to low values of tenure and not having a two-year contract.
- High values of PC2 corresponds to having InternetService_Fiber.optic, and higher tenure.

Decision Boundary and Interpretation

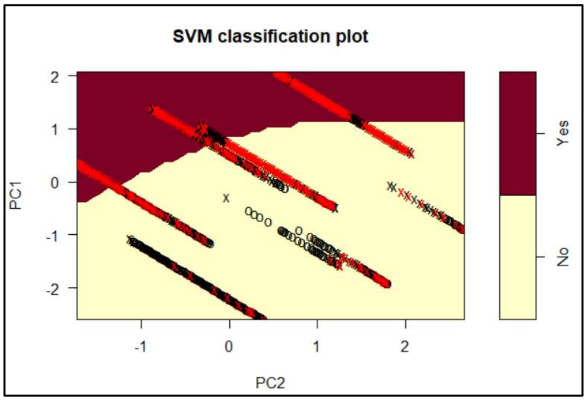


Figure 10 : SVM classification plot

Figure 10 plots the SVM decision boundary, support vectors (x-marks), and training data (o-marks) in two dimensions.

From the plot, it is clear that high values of PC1 and low values of PC2 correspond to churn. In terms of the original features, this suggests that the customers most likely to churn are:

- Newer customers (lower tenure)
- Customers on one year/month to month contracts
- Customers using DSL internet or not subscribing to the internet service.

Model Comparison

We compared the two models based on the following parameters:

- **Performance**

  The SVM model performed better on the test data than the Decision Tree model. Both models fit the data well – there is very little difference between their performance on the training and test data.

- **Interpretability**

  The results of the decision tree model easier to interpret and don't need any additional processing to obtain. To obtain interpretable decision boundary for the SVM model we had to use PCA to transform the data and had to first understand how the features impacted the principal components and then use this knowledge to interpret the SVM plot.

- **Computational Complexity**
  Due to the added step of PCA transforming variables, the SVM model was computationally more complex. Additionally the standard SVM model took longer to run than the Decision Tree model.

Recommendations

Based on our analysis of the dataset we would make the following recommendations to the company:

1. Incentivize customers to switch to longer-term plans.
2. Customers who use the company's fiber optic internet service as well are less likely to churn. Lower cost first-year plans that encourage adoption should be pitched to customers. Additionally, customers on the DSL plan should be encouraged to switch to an optic fiber connection where possible.
3. Encourage customers to switch to automated billing via credit card or bank transfer. Customers making payments manually each month using methods like echecks have a higher churn rate.

Limitations

- We did not tune the SVM model. Tuning the SVM model could have potentially further improved the results. ~~were unable to tune the SVM model after adding the class weigh~~

References

Dataset:
https://www.kaggle.com/blastchar/telco-customer-churn

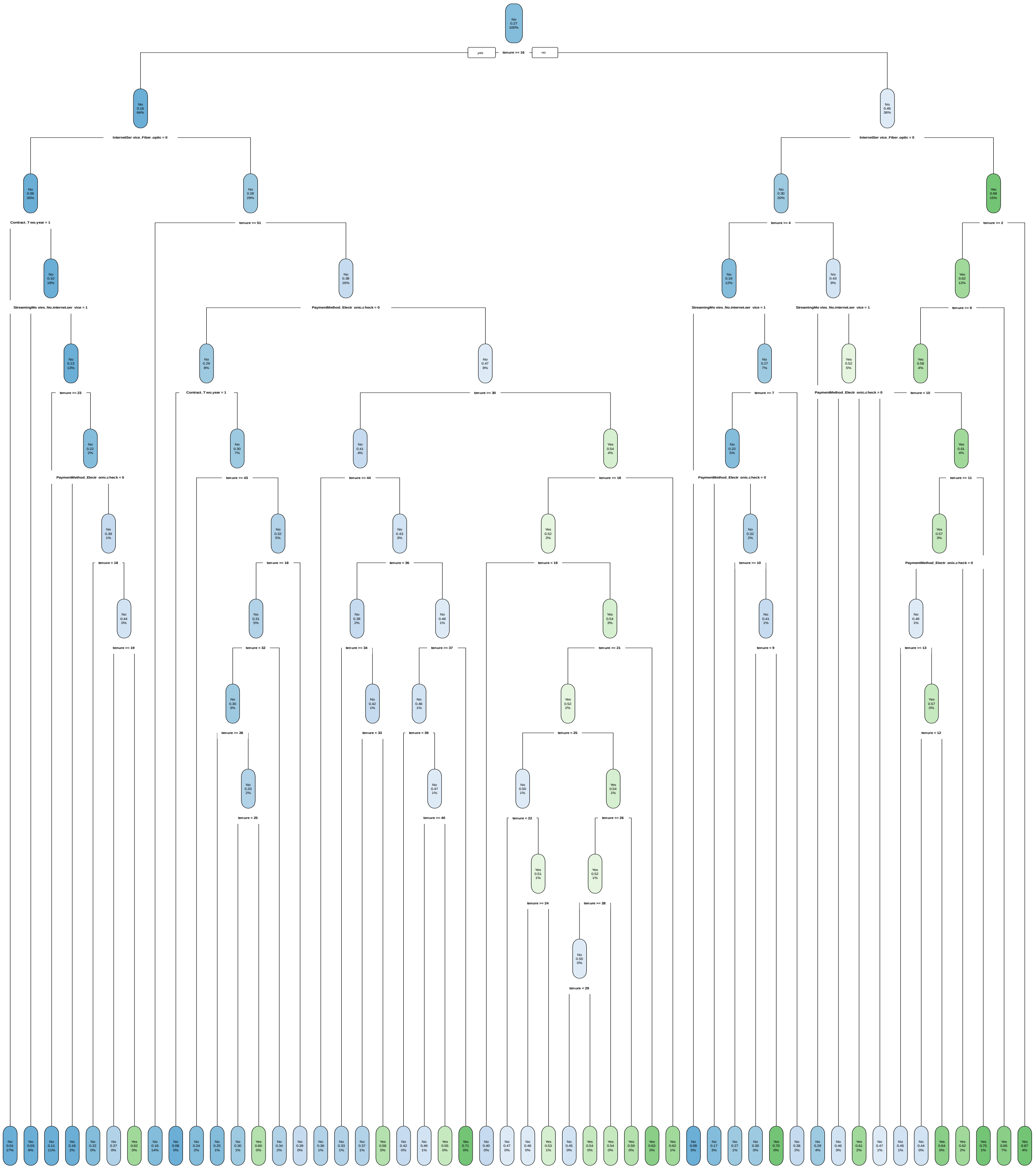
Exhibit 1: Data Overview

Variables	**Type**	**Cardinality or Range**	**Preprocessing Act**
customerID	Categorical	Distinctive	Delete
gender	Categorical	2 (Female/Male)	To 1 dummy variable
SeniorCitizen	Categorical	2 (0/1)	
Partner	Categorical	2 (Yes/No)	To 1 dummy variable
Dependents	Categorical	2 (Yes/No)	To 1 dummy variable
tenure	Numerical	0-72 (Integer)	
PhoneService	Categorical	2 (Yes/No)	To 1 dummy variable
MultipleLines	Categorical	3 (Yes/No/No phone service)	To 2 dummy variables
InternetService	Categorical	3 (DSL/Fiber optic/No)	To 2 dummy variables
OnlineSecurity	Categorical	3 (Yes/No/No Internet service)	To 2 dummy variables
OnlineBackup	Categorical	3 (Yes/No/No Internet service)	To 2 dummy variables
DeviceProtection	Categorical	3 (Yes/No/No Internet service)	To 2 dummy variables
TechSupport	Categorical	3 (Yes/No/No Internet service)	To 2 dummy variables
StreamingTV	Categorical	3 (Yes/No/No Internet service)	To 2 dummy variables
StreamingMovies	Categorical	3 (Yes/No/No Internet service)	To 2 dummy variables
Contract	Categorical	3 (One year/Two year/M-to-M)	To 2 dummy variables
PaperlessBilling	Categorical	2 (Yes/No)	To 1 dummy variable
PaymentMethod	Categorical	4 (Bank (Auto)/Credit C (Au…)	To 3 dummy variables
MonthlyCharges	Numerical	18.25-118.75	
TotalCharges	Numerical	18.8-8684.8	
Churn	Categorical	2 (Yes/No)	To 1 dummy variable

Exhibit 2: Removed Variables in High Correlation Pairs

MonthlyCharges	InternetService_No	OnlineSecurity_No.internet.service
OnlineBackup_No.internet.service	DeviceProtection_No.internet.service	TechSupport_No.internet.service
StreamingTV_No.internet.service	TotalCharges	MultipleLines_No.phone.service

Exhibit 3: Decision Tree Plot (Model 2)



No 0.27 100%
yes
tenure >= 16
no
No 0.16 64%
No 0.46 36%
InternetSer vice_Fiber .optic = 0
InternetSer vice_Fiber .optic = 0
No 0.06 35%
No 0.28 29%
No 0.30 20%
Yes 0.68 15%
Contract_T wo.year = 1
tenure >= 51
tenure >= 4
tenure >= 2
No 0.10 19%
No 0.38 16%
No 0.19 12%
No 0.43 9%
Yes 0.62 12%
StreamingMo vies_No.internet.ser vice = 1
PaymentMethod_Electr onic.check = 0
StreamingMo vies_No.internet.ser vice = 1
StreamingMo vies_No.internet.ser vice = 1
tenure >= 9
No 0.13 13%
No 0.29 8%
No 0.47 8%
No 0.27 7%
Yes 0.52 5%
Yes 0.58 4%
tenure >= 23
Contract_T wo.year = 1
tenure >= 30
tenure >= 7
PaymentMethod_Electr onic.check = 0
tenure < 10
No 0.22 2%
No 0.30 7%
No 0.41 4%
Yes 0.54 4%
No 0.22 5%
Yes 0.61 4%
PaymentMethod_Electr onic.check = 0
tenure >= 43
tenure >= 44
tenure >= 18
PaymentMethod_Electr onic.check = 0
tenure >= 11
No 0.39 1%
No 0.32 5%
No 0.43 3%
Yes 0.52 3%
No 0.32 2%
Yes 0.57 3%
tenure < 18
tenure >= 18
tenure < 36
tenure < 19
tenure >= 10
PaymentMethod_Electr onic.check = 0
No 0.44 0%
No 0.31 5%
No 0.39 2%
No 0.48 1%
Yes 0.54 3%
No 0.41 1%
No 0.49 1%
tenure >= 19
tenure < 32
tenure >= 34
tenure >= 37
tenure >= 21
tenure < 9
tenure >= 13
No 0.30 3%
No 0.42 1%
No 0.46 1%
Yes 0.52 2%
Yes 0.57 0%
tenure >= 26
tenure < 33
tenure < 39
tenure < 25
tenure < 12
No 0.33 2%
No 0.47 1%
No 0.50 1%
Yes 0.54 1%
tenure < 25
tenure >= 40
tenure < 22
tenure >= 26
Yes 0.51 1%
Yes 0.52 1%
tenure >= 24
tenure >= 28
No 0.50 0%
tenure < 29
No 0.01 17%
No 0.03 6%
No 0.11 11%
No 0.16 2%
No 0.22 0%
No 0.37 0%
Yes 0.62 0%
No 0.16 14%
No 0.08 0%
No 0.24 2%
No 0.25 1%
No 0.30 1%
Yes 0.60 0%
No 0.34 2%
No 0.39 0%
No 0.36 1%
No 0.33 1%
No 0.37 1%
Yes 0.58 0%
No 0.42 0%
No 0.46 1%
Yes 0.55 0%
Yes 0.71 0%
No 0.40 0%
No 0.47 0%
No 0.48 0%
Yes 0.53 1%
No 0.45 0%
Yes 0.54 0%
Yes 0.54 0%
Yes 0.59 0%
Yes 0.63 0%
Yes 0.62 1%
No 0.08 5%
No 0.17 3%
No 0.27 1%
No 0.30 0%
Yes 0.70 0%
No 0.38 2%
No 0.29 4%
No 0.46 3%
Yes 0.61 2%
No 0.47 1%
No 0.45 1%
No 0.44 0%
Yes 0.64 0%
Yes 0.62 2%
Yes 0.75 1%
Yes 0.65 7%
Yes 0.87 4%