Gramener Customer Churn Modeling Challenge

Stage 2 Submission

Vishnu

Churn Prediction Challenge

- Churn Business Problem
  - Churn represents the loss of an existing customer to a competitor
  - Churn is a problem for any provider of a subscription service or recurring purchasable
- Predicting churn is the key to a protective strategy
  - Can assist churn management by tagging customers most likely to churn
  - High risk customers should first be sorted by profitability
  - Campaign targeted to the most porfitable at-risk customers
  - Retention campaigns must be targeted to the right customers
- Challenge Problem
  - Given information about customers of a telecom operator
  - Identify the customers most likely to defect

Nature of Data and Preprocessing

- Data were provided for 100,000 customers
  - Complete dataset - no attributes missing for a customer
  - Highly imbalanced dataset: contains 2292 non-churners and 375 churners
- Feature selection
  - Target variable: Churn
  - Categorical variables included: International plan, Message Plan
  - Continuous variables included: Day Calls, Day Mins, Eve Calls, Eve Mins, Night Calls, Night Mins, International Calls, International Mins, Account Length, Messages, CustomerService Calls
- Features removed:
  - 4 Charge variables which are linear multiples of 4 Mins variables were eliminated
  - Area Code, since it contains only 3 different values
  - State contains 51 levels, with insufficient information in each level
  - Phone doesn't contain relevant data that can be used for prediction

Preprocessing

- Newly created features:
  - Total Calls, Total Mins, Total Charge
  - 4 *PropMins variables, measures proportion of call duration.
    eg: Day PropMins = Day Mins / Total Mins
  - 4 *PropCalls variables, measures proportion of number of calls made.
    eg: Day PropCalls = Day Calls / Total Calls
  - MessagesPerWeek = Messages / Account Length
  - 4 *AverageMinsPerCall variables, eg: Day AverageMinsPerCall = Day Mins / Day Calls
- Minimize feature redundancy:
  - Features having very low variance (few unique values) were removed
  - Highly correlated variables were eliminated
  - Boruta algorithm was used to select the final set of features
- To handle the imbalance in the data new synthetic churners were created using SMOTE sampling technique

Predictive Modeling (Strategy)

- Several learning algorithms were trained on the processed dataset
    - GLMs, Tree based models, SVMs, Neural networks, Vowpal Wabbit
- 5 fold repeated cross validation was used to tune model parameters
- Data was split into 70/30 (train/test) and base models were selected using their performance on the test.
- F1 metric was used to measure the performance of models.

$$F1 = 2\,\frac{pr}{p+r} \quad \text{where} \;\; p = \frac{tp}{tp+fp}\,, \quad r = \frac{tp}{tp+fn}$$

- Measures accuracy using the statistics precision p and recall/sensitivity r.
- Particularly useful in imbalanced datasets where the cost of misclassification of a positive is higher than misclassifying a negative

Predictive Modeling (Comparison of base models)

MODEL	F1 SCORE (OOS)
C5.0	0.9209476
Adaboost	0.9010143
Adacost	0.9126099
Random Forest	0.9082373
SVM (Radial)	0.7161835
Bagged Tree	0.8866876
GBM	0.8939888
GLM	0.4295080
GAM	0.5189876
Neural network (nnet)	0.7216470
Oblique RF (logistic)	0.6980502

Predictive Modeling (contd)

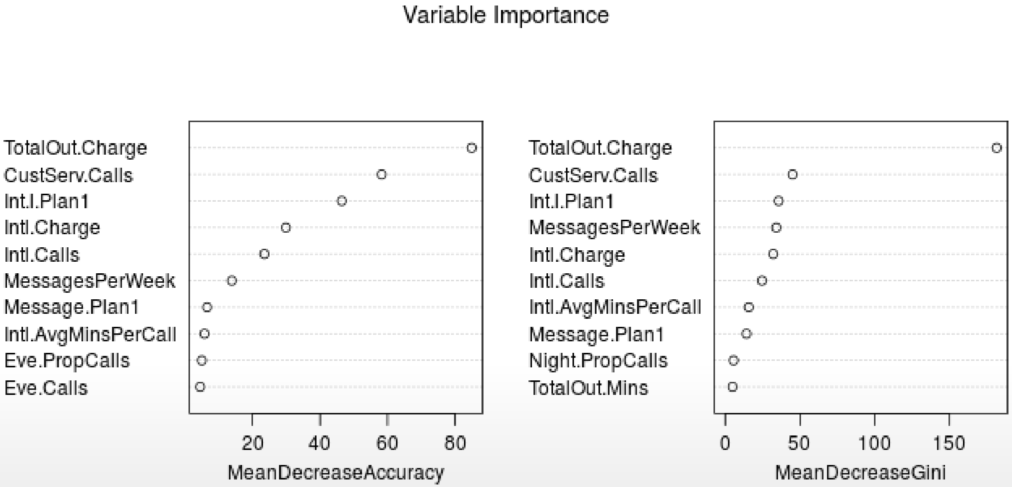
- Classification trees (rpart model) had very low sensitivity (CV Sensitivity $= 0.3870$) which implies greater number of false non-churners
- Cost sensitive boosted tree models were found to give the best results (F1 score ≥ 0.9)
- Bagged tree models were the next best perfomers (F1 score ≈ 0.9)
- Regularization applied to random forests or linear models didn't improve performance.
- Vowpal Wabbit gave a poor generalization error (F1 score ≤ 0.5), and hence was discarded.
- Tools Used:
  - R 3.2.2
  - Python 2.7
  - Vowpal Wabbit 8.0.0

Final Prediction Model

- Stack of multiple tree models. For every bagged model another complementary model was trained using SMOTE sampled data to reduce the number of false negatives, thereby stabilizing the model and increasing sensitivity.
  - Adacost
  - Random Forest
  - cost sensitive C5.0 tree
  - xgboost implementation of GBM
  - a model ensemble of Bagged Trees
- Base models were stacked using an aggregate of several neural networks
- Final model was trained using the entire dataset

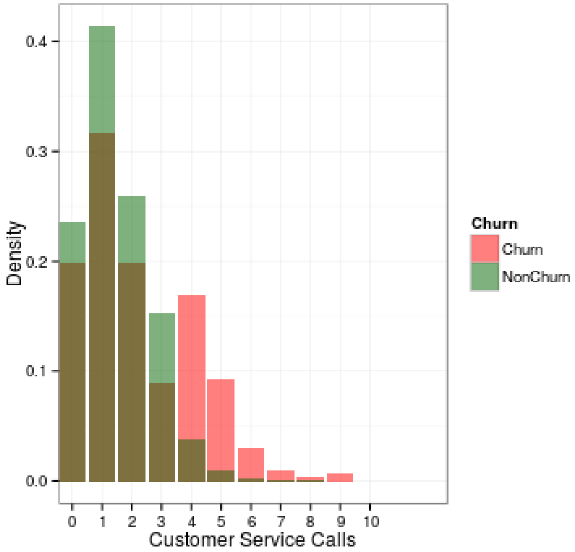
Analysis of Results - Variable Importance

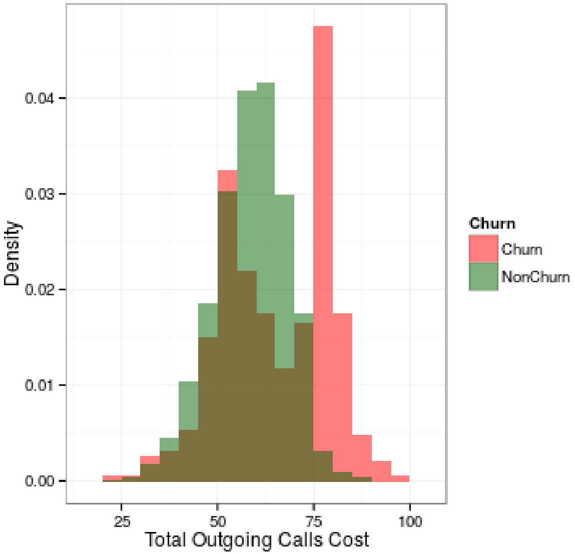
- Assessing variable importance guides in finding the most influential features affecting the propensity of risk to churn
- Random forests were well performing on the dataset and hence was used to determine the variable importance
- Density plots were used to interpret the results. Preferred over conventional frequency histograms due to imbalance in the dataset

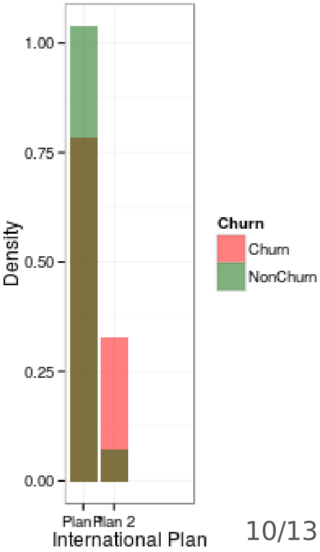


Analysis of Results - Influential Variables

- The following categories of customers have greater probability that they belong to the population of churners than the non-churners
  - Customers who make 4 or more calls to the customer service
  - Customers whose total outgoing calls cost is greater than 75$
  - Customers who are subscribers of international calls Plan 2

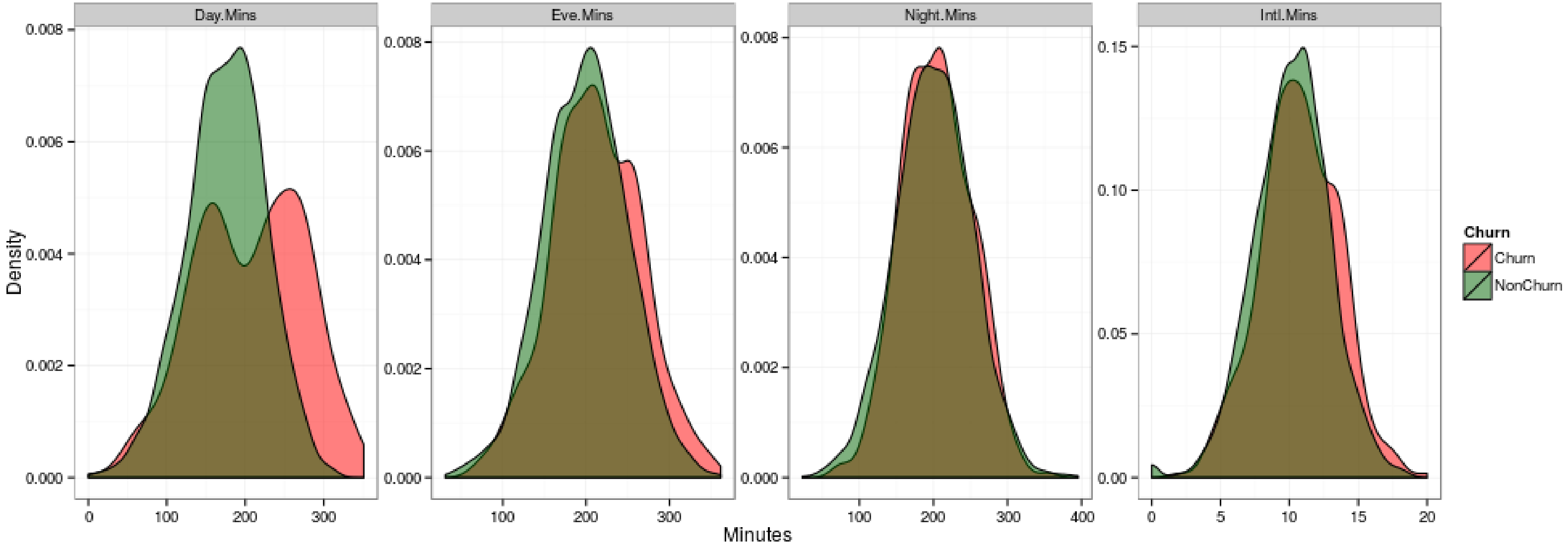






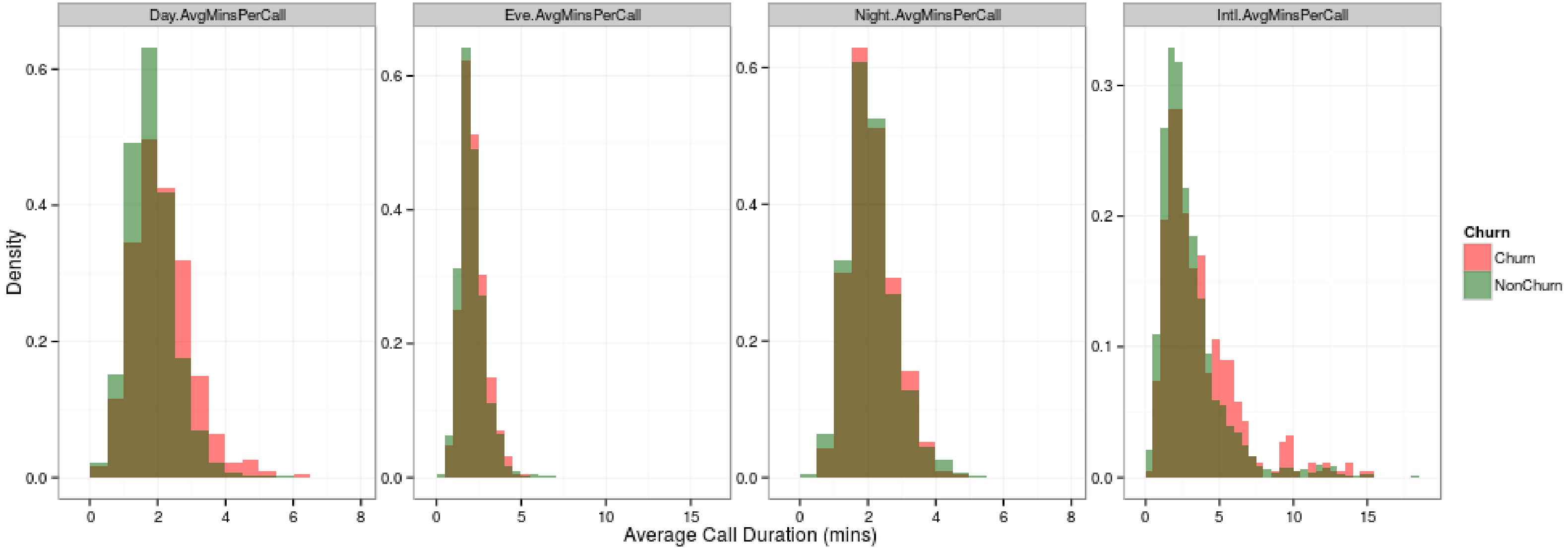
Call Duration Variables

- Customers whose total duration of daytime calls exceed 275mins have greater probability that they belong to the population of churners, and hence have high risk to churn



Average Call Duration

- Customers whose average duration of daytime calls exceed 2.5mins are more likely to churn. Alongwith the previous result, this is a strong indication that customers are unsatisfied with the daytime call tariffs.



Conclusion

- Not all the variables have significance in predicting churn
- Total Charge, CustomerService Calls, International Plan, Call Duration variables were effectual in assessing the risk of churn
- A successful model for prediction and prevention of churn in telecommunication companies can influence very positively an overall profit of companies
- Prediction model helps to combat churn by identifying customers most likely to defect and taking preventative measures (offering incentives etc) with customers you want to keep