CUSTOMER CHURN

HOW TO IDENTIFY AND PREVENT CUSTOMER LOSS

Sara Herbstreit
DSC630

THE PROBLEM

- Customer churn is a major problem in the telecommunication (telecom) industry.
- Though there are many reasons a customer may leave a service, many of them are preventable.
- Customer churn in the telecom industry costs 780 million dollars for U.S. carriers annually (Aditya Kapoor, 2017).

THE SOLUTION

- Identify customers that are at high risk for leaving and implement mitigation techniques to reduce the churn rate.
- To achieve the reduction in churn rate, a historical telecom dataset was obtained from Kaggle.com, containing information on more than 3,000 customers.
- This data will be used to train machine learning models, which will learn to predict when a customer is about to leave
- Once a customer has been identified as high churn risk, mitigation processes can begin to retain this customer.

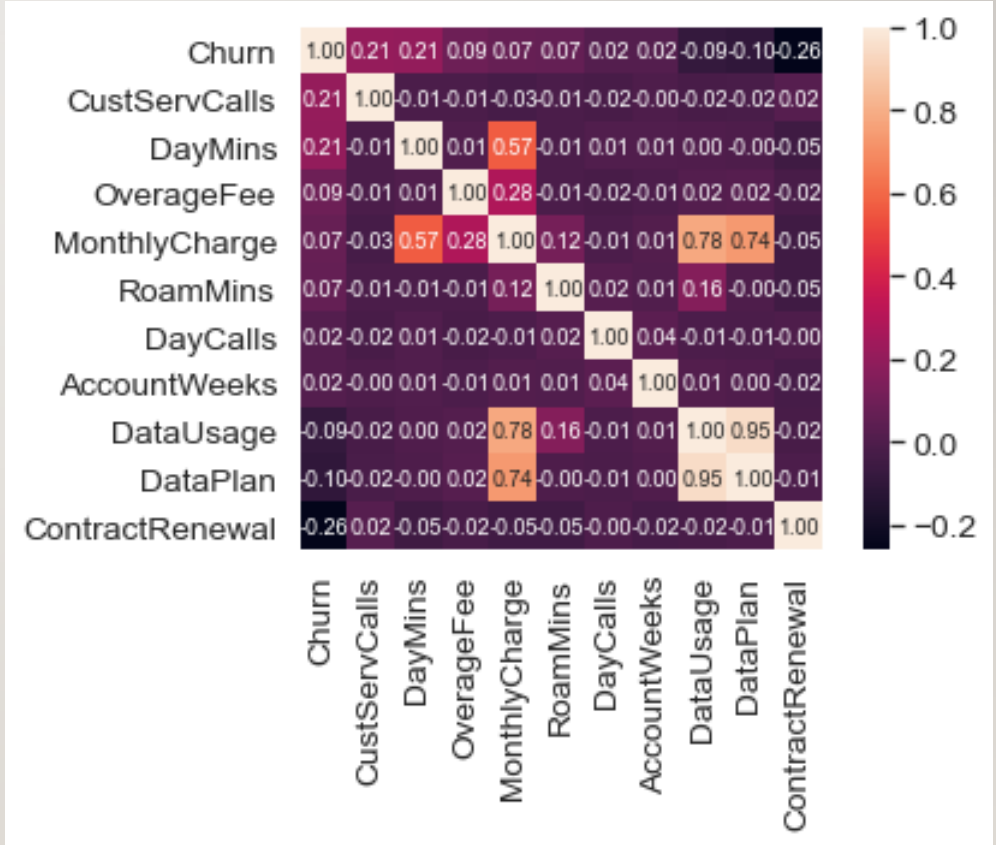
THE DATA

- The data contains 10 independent variables, with the target feature being "Churn"
- The starting data set has an imbalanced target class with only 482/3333 points representing customers that left.
- Outliers were identified, but due to them being predominantly in the minority of the target class (composed 33% of this class), they were deemed as important and left in

Churn	1 if customer cancelled service, 0 if they did not
AccountWeeks	number of weeks customer had an active account
ContractRenewal	1 if customer recently renewed contract, 0 if they did not
DataPlan	1 if customer has a data plan, 0 if they do not
DataUsage	gigabytes of monthly data usage
CustServCalls	number of calls into customer service
DayMins	average daytime minutes per month
DayCalls	average number of daytime calls
MonthlyCharge	average monthly bill
OverageFee	largest overage fee in last 12 months
RoamMins	average number of roaming minutes

RELATIONSHIPS TO TARGET

- A heatmap was created to look at potential relationships to the target variable.
- We can quickly see that customer service calls, daytime minutes used, and contract removal have the strongest relationships to the target feature



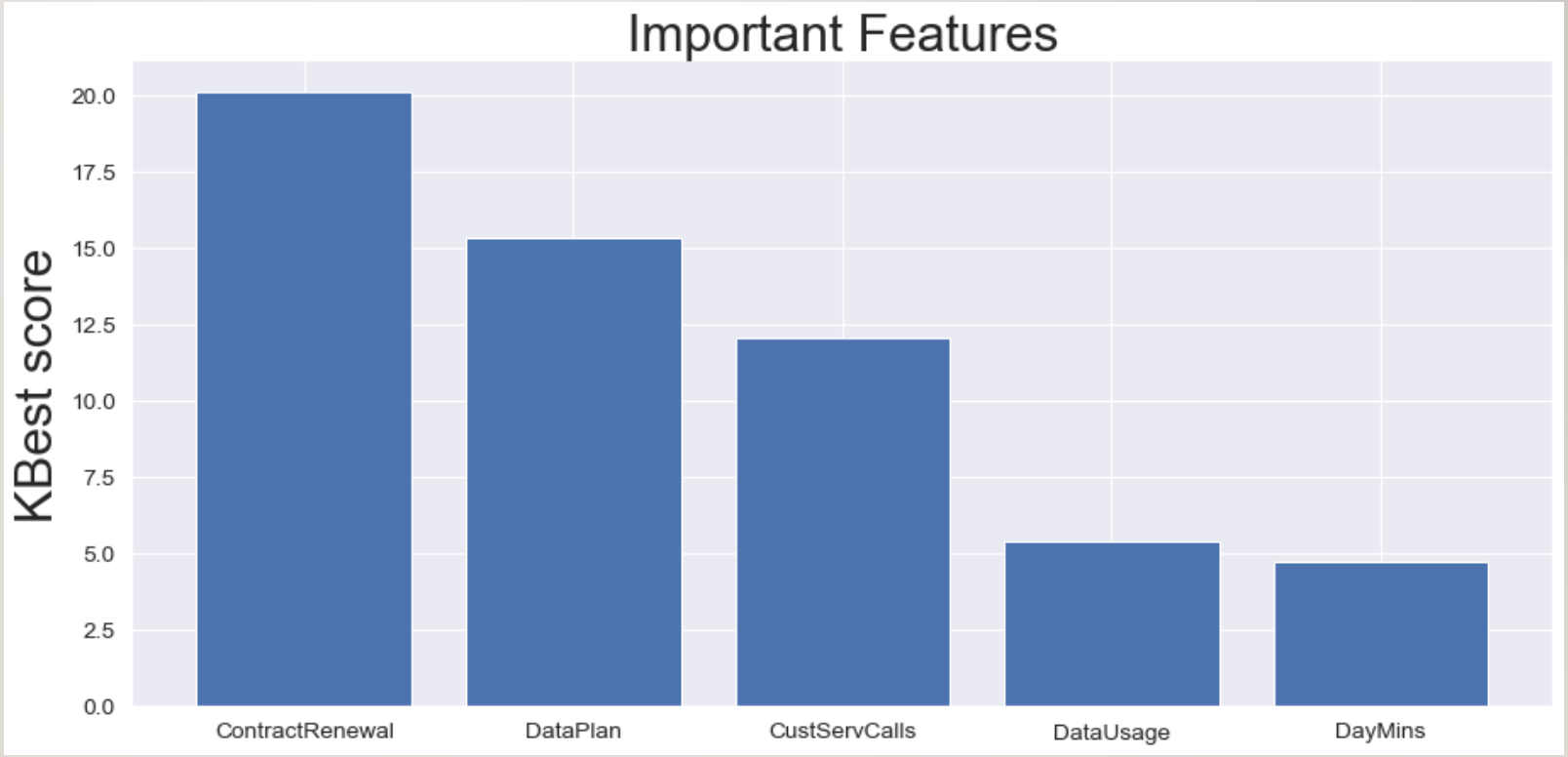
PREPROCESSING AND MODEL SELECTION

- The data was split into training and test sets prior to any transformations.
- The target minority class was upsampled in the training set to match the size of the majority using SMOTE.
- The training features were scaled so the range of all values was 0-1.
- The tranformed data was used to train 6 models
- The Random Forest model beat other models with a ROC AUC score of 0.87.
- F1-score was caluculated on the Random Forest model. The majority class f1 score was 0.97, and minority class was 0.81.

Model	ROC AUC Score
LogisticRegression	0.809
SVC	0.873
LinearSVC	0.811
KNeighbors	0.824
DecisionTree	0.778
RandomForest	0.879

FEATURE SELECTION

- Feature selection using the chi2 statistic was used to identify the top 5 features in the data set
- The Random Forest model was re-trained using only the top 5 features. This resulted in a slight decrease from the baseline score. The full selection of features was chosen to move forward.

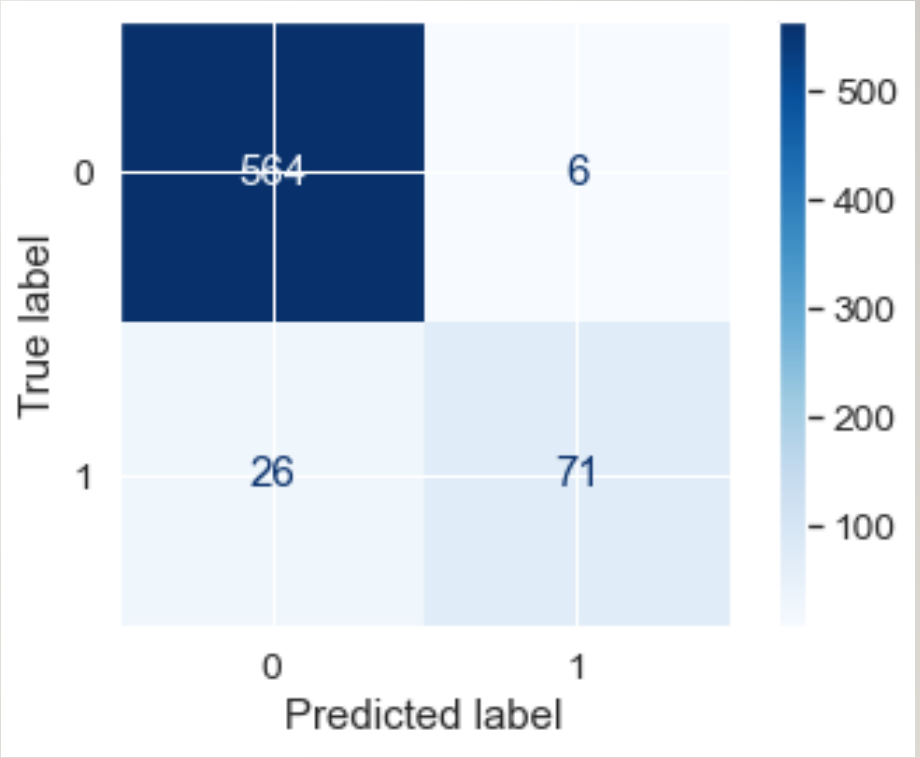


HYPERTUNING THE MODEL

- A grid search was performed to find the best possible combination of hyperparameter adjustments. The following hyperparameters were chosen:
  - n_estimators = 522
  - max_features = sqrt
  - max_depth = 110
  - min_samples_split = 2
  - min_samples_leaf = 2
  - Bootstrap = True
- The Random Forest model was re-trained using these
- The final model has a performance score of 0.901 on the test data.

RESULTS

- The confusion matrix shows that the model is excellent at labelling the non-churn class and can identify the churn class approximately 73.2% of the time.
- This model could be used to identify customers about to churn, with an extremely low false positive rate. This makes it great to minimize the loss of resources on customers that aren't going to churn.



MODEL IMPACT

- According to a survey on statista.com, 21% of respondants switched to a new telecom service provider in the last year
- If mitigation is 100% successful, this rate would be as low as 5.6%
- This model presents very little risk to wasting resources