Predicting Customer Churn

DATA621 – Final Project

BY: ERIC LEHMPHUL

Why focus on detecting customer churn?

- Businesses rely on customers to purchase products or services from them to supply revenue and create profits.
- Customers leaving a business leads to:
  - Higher customer acquisition cost
  - Reduced business revenue

Customer Churn – Related Literature

- Class Imbalance
  - Ali et al. (2019) has expressed three techniques to handle the imbalance present in the target variable: External, Algorithmic / internal, and Cost-sensitive.
- Modeling Approaches
  - Dahiya & Bhatia (2015) used logistic regression and decision trees
  - Almana et al. (2014) used neural networks, decision trees, covering algorithms, and statistical models
  - Yi Fei et al. (2017) applied Naïve Bayes and K-means

Customer Churn Dataset

- Contains 4250 observations and 20 columns
- Composed of
  - Metadata related to the customer like activity levels and geographic location
- Link to dataset: https://www.kaggle.com/competitions/customer-churn-prediction-2020/data?select=train.csv

Variable	Data Type	Description
state	String	2-letter code of the US state of customer residence
account_length	Numeric	Number of months the customer has been with the current telco provider
area_code	String	3 digit area code of customer
international_plan	String	Indicator variable to identify if customer has an international plan
voice_mail_plan	String	Indicator variable to identify if customer has a voice mail plan
number_vmail_messages	Numeric	Number of voice mail messages recieved by customer
total_day_minutes	Numeric	Total minutes of day calls
total_day_calls	Numeric	Total number of day calls
total_day_charge	Numeric	Total charge of day calls
total_eve_minutes	Numeric	Total minutes of evening calls
total_eve_calls	Numeric	Total number of evening calls
total_eve_charge	Numeric	Total charge of evening calls
total_night_minutes	Numeric	Total minutes of night calls
total_night_calls	Numeric	Total number of night calls
total_night_charge	Numeric	Total charge of night calls
total_intl_minutes	Numeric	Total minutes of international calls
total_intl_calls	Numeric	Total number of international calls
total_intl_charge	Numeric	Total charge of international calls
number_customer_service_calls	Numeric	Number of calls to customer service
churn	String	Binary indicator variable to identify if customer churned

Methodology

Conduct

- Conduct external class data balancing by creating an undersampled dataset and an oversampled dataset.

Create

- Create logistic regression and naïve bayes models to predict customer churn using the three dataset: original data, undersampled data, and oversampled data.

Assess

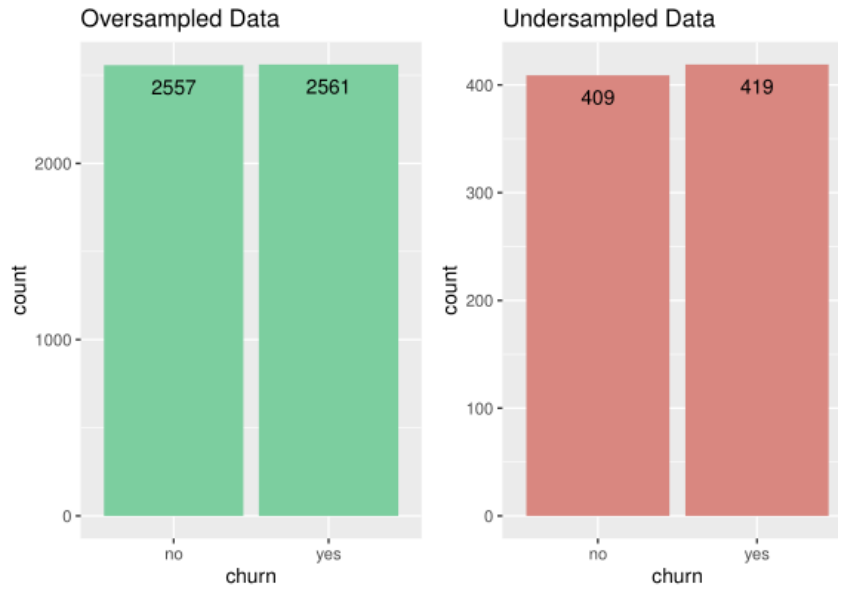
- Assess model performance via confusion matrix metrics, especially precision

Summary Statistics

- There are no missing values present in the data
- 5 variables are factor variables and the remaining 15 are numeric
- There is a large class imbalance in the target variable churn (no = 3652, yes = 598

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
state	0	1	FALSE	51	WV: 139, MN: 108, ID: 106, AL: 101
area_code	0	1	FALSE	3	415: 2108, 408: 1086, 510: 1056
international_plan	0	1	FALSE	2	no: 3854, yes: 396
voice_mail_plan	0	1	FALSE	2	no: 3138, yes: 1112
churn	0	1	FALSE	2	no: 3652, yes: 598

skim_variable	n_missing	complete_rate	mean	sd	median	min	max
account_length	0	1	100.24	39.70	100.00	1	243.00
number_vmail_messages	0	1	7.63	13.44	0.00	0	52.00
total_day_minutes	0	1	180.26	54.01	180.45	0	351.50
total_day_calls	0	1	99.91	19.85	100.00	0	165.00
total_day_charge	0	1	30.64	9.18	30.68	0	59.76
total_eve_minutes	0	1	200.17	50.25	200.70	0	359.30
total_eve_calls	0	1	100.18	19.91	100.00	0	170.00
total_eve_charge	0	1	17.02	4.27	17.06	0	30.54
total_night_minutes	0	1	200.53	50.35	200.45	0	395.00
total_night_calls	0	1	99.84	20.09	100.00	0	175.00
total_night_charge	0	1	9.02	2.27	9.02	0	17.77
total_intl_minutes	0	1	10.26	2.76	10.30	0	20.00
total_intl_calls	0	1	4.43	2.46	4.00	0	20.00
total_intl_charge	0	1	2.77	0.75	2.78	0	5.40
number_customer_service_calls	0	1	1.56	1.31	1.00	0	9.00



Balancing Data

- Undersampling and Oversampling were used
  - Undersampling reduces the number of the majority class to match minority class via random sampling
  - Oversampling increases the number of the minority class to match majority class via random sampling and or synthesizing data

Logistic Regression Models

- Background:
  - Regression modeling technique that can find the probability of a binary variable and classify the probability given a threshold
  - The model creates coefficients that allow for the model to be interpreted
- Models:
  - Model 1: Unbalanced Data Logistic Regression
  - Model 2: Oversampled Data Logistic Regression
  - Model 3: Undersampled Data Logistic Regression

	Model 1	Model 2	Model 3
Accuracy	0.868	0.775	0.751
Sensitivity	0.577	0.355	0.335
Specificity	0.885	0.948	0.954
Pos.Pred.Value	0.229	0.737	0.782
Neg.Pred.Value	0.973	0.781	0.746
Precision	0.229	0.737	0.782
Recall	0.577	0.355	0.335
F1	0.328	0.479	0.469
Prevalence	0.056	0.292	0.328
Detection.Rate	0.032	0.104	0.110
Detection.Prevalence	0.141	0.141	0.141
Balanced.Accuracy	0.731	0.651	0.645
AUC	0.812	0.816	0.814
AUC_PR	0.476	0.461	0.450

- Findings:
  - The unbalanced data (Model 1) was not effective at detecting customer churn
  - The oversampled data (Model 2) was produced the best F1 score, indicating the highest combination of precision and recall
  - The undersampled data (Model 3) was the best at identifying a customer that has churned, as it had the highest precision score

Naïve Bayes Models

- Background:
  - Modeling technique that assumes independence among variables
  - It produces probabilities for each class and assigns the data point to the class with the highest probability
- Models:
  - Model 4: Unbalanced Data Naïve Bayes
  - Model 5: Oversampled Data Naïve Bayes
  - Model 6: Undersampled Data Naïve Bayes

	Model 4	Model 5	Model 6
Accuracy	0.881	0.756	0.740
Sensitivity	0.696	0.336	0.322
Specificity	0.891	0.950	0.952
Pos.Pred.Value	0.268	0.754	0.771
Neg.Pred.Value	0.981	0.756	0.735
Precision	0.268	0.754	0.771
Recall	0.696	0.336	0.322
F1	0.387	0.465	0.455
Prevalence	0.054	0.316	0.336
Detection.Rate	0.038	0.106	0.108
Detection.Prevalence	0.141	0.141	0.141
Balanced.Accuracy	0.793	0.643	0.637
AUC	0.862	0.799	0.803
AUC_PR	0.538	0.505	0.499

- Findings:
  - The unbalanced data (Model 4) was not was not great at predicting customer churn, lowest precision
  - Model 5 produced the highest F1 score
  - Model 6 is the best Naïve Bayes model at predicting customer churn, as the precision is the highest.

Comparing Logistic Regression and Naïve Bayes Models

- The logistic regression and naive bayes models performed similarly across the 6 models
- Logistic Regression outperformed the naive bayes classifier when using undersampled data (Model 3 vs Model 6) in terms of precision, recall, f1 score, and detection rate
- The Naive Bayes classifier was a better model when using oversampled data (Model 2 vs Model 5) because it had a higher precision, recall, f1 score, and detection rate
- The best model is Model 3 - Logistic Regression with undersampled data because it yields the highest precision in classifying that a customer will churn.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Accuracy	0.868	0.775	0.751	0.881	0.756	0.740
Sensitivity	0.577	0.355	0.335	0.696	0.336	0.322
Specificity	0.885	0.948	0.954	0.891	0.950	0.952
Pos.Pred.Value	0.229	0.737	0.782	0.268	0.754	0.771
Neg.Pred.Value	0.973	0.781	0.746	0.981	0.756	0.735
Precision	0.229	0.737	0.782	0.268	0.754	0.771
Recall	0.577	0.355	0.335	0.696	0.336	0.322
F1	0.328	0.479	0.469	0.387	0.465	0.455
Prevalence	0.056	0.292	0.328	0.054	0.316	0.336
Detection.Rate	0.032	0.104	0.110	0.038	0.106	0.108
Detection.Prevalence	0.141	0.141	0.141	0.141	0.141	0.141
Balanced.Accuracy	0.731	0.651	0.645	0.793	0.643	0.637
AUC	0.812	0.816	0.814	0.862	0.799	0.803
AUC_PR	0.476	0.461	0.450	0.538	0.505	0.499

Conclusion

- The best model for identifying the positive class, churn = yes, was Model 3, a logistic regression model trained with the undersampled dataset.
- Balancing the data produced much better results at classifying a customer that is likely to leave the business. The logistic regression and naïve bayes precision for the unbalanced model was 0.229 and 0.268 respectively. The balanced models produced far better results with precisions equal to 0.737 and 0.782 for logistic regression and 0.754 and 0.771 for naïve bayes.
- The resulting models created in this study allow for customer churn to be detected with high precision. Given the same input features, a telecommunication company can detect potential customers of interest that are likely to leave. The telecommunication company can choose to either accept that the consumer will churn or they can take steps to retain the customer for continued service.

References

- Ali, H., Nahib Mohd Salleh, M., Saedudin, R., Hussain, K., & Mushtaq, M. F. (2019). Imbalance classproblems in data mining: a review. Indonesian Journal of Electrical Engineering and Computer Science, 14(3), 1560–1571. https://doi.org/10.11591/ijeecs.v14.i3.pp1560-1571
- Almana, A. M., Aksoy, M. S., & Alzahran, R. (2014). A Survey On Data Mining Techniques In Customer
- Churn Analysis For Telecom Industry. Int. Journal of Engineering Research and Applications, 4(5), 165–171. Retrieved from https://www.ijera.com/papers/Vol4_issue5/Version%206/AF4506165171.pdf. 20
- Dahiya, K., & Bhatia, S. (2015). Customer churn analysis in telecom industry. IEEE. Retrieved from https://ieeexplore.ieee.org/abstract/document/7359318/authors#authors.
- Yi Fei, T., Hai Shuan, L., Jie Yan, L., Xiaoning, G., & Wooi King, S. (2017). Prediction on Customer Churn in the Telecommunications Sector Using Discretization and Naïve Bayes Classifier. Int. Journal Advanced Software Computer. Applications, 9(3), 24–35. Retrieved from https://www.i-csrs.org/Volumes/ijasca/2_Page-23_35_Predictive-Analysis-for-Telecommunications-Customer-Churn-on-Big-Data-Platform.pdf.