PREDICTING HOSPITAL READMISSIONS TO PREVENT THEM

JOLENE BRANCH, ANDREA FOX, & SAM SEARS
DSC630 PREDICTIVE ANALYTICS, BELLEVUE UNIVERSITY

BACKGROUND OF THE PROBLEM

Before hospitalization:

Inadequate social and medical resources

After hospitalization:

Suboptimal care transitions

Inadequate social and medical resources

Vicious cycle of readmission

BUSINESS UNDERSTANDING FOR PREVENTING READMISSIONS

- Excessive cost to patient and third-party payers
- Bed capacity crises
  - Additional construction
  - Need to prioritize or restrict care
- Reduces future Medicare payments to the facility
- Weakens relationship with community

DATA UNDERSTANDING

- UCI Machine Learning Repository – readmissions for diabetic patients (Beata, 2014)
  - 1999-2008
  - 130 hospitals
  - 50+ features
    - Categorical
    - Target variable = readmitted
- Mean readmission rate = 11.2%

DATA PREPARATION - INITIAL

- Removal of hospice patients and deaths
- Kept top 10 medical specialties and converted less frequent values to 'other' due to skewed distribution
- Filling Nan values with 'UNK' (unknown)
- One-hot encoding of categorical variables to be able to use sklearn algorithms
- Binned age in years to intervals of 10
- Converting weight variable to binary 'has_weight' due to lots of missing values
- Created **binary target feature** using a derived variable based on "Readmission within 30 days - Yes/No?"

INITIAL MODELING

- Split data into three sets: 15% test, 15% validation, and 70% training
- Balanced training set to have 50% positive and 50% negative to account for an uneven distribution
- Scaled training and validation data inputs using sklearn StandardScaler()
- Training data fit and validation data tested on 7 different classification models and a heterogeneous ensemble of all models
  - K-Nearest Neighbor, Logistic Regression, Stochastic Gradient Descent, Naïve Bayes, Decision Tree, Random Forest, and Gradient Boosting

INITIAL RESULTS

- Compared AUC, Accuracy, Precision, and Recall scores between train and validation for generalization and between models for performance
- Overall, scores were similar between both except for Precision, as expected, because the train data set was balanced, and the validation data set was not.

	Model	Train AUC	Train Accuracy	Train Precision	Train Recall	Validation AUC	Validation Accuracy	Validation Precision	Validation Recall
7	Ensemble	0.641577	0.641577	0.664761	0.571220	0.621363	0.666109	0.189686	0.562871
2	SGD	0.615492	0.615492	0.632987	0.549715	0.618973	0.663252	0.187773	0.561091
1	LogReg	0.620935	0.620935	0.639211	0.555290	0.618419	0.663182	0.187488	0.559905
5	RF	0.638856	0.638856	0.647117	0.610779	0.618146	0.637745	0.181307	0.592527
6	GBC	0.696071	0.696071	0.704741	0.674897	0.602600	0.619834	0.170830	0.580071
4	DTC	0.666999	0.666999	0.693717	0.598035	0.600352	0.654889	0.176634	0.529063
0	KNN	0.600358	0.600358	0.621778	0.512412	0.583587	0.650707	0.167267	0.495848
3	NB	0.503120	0.503120	0.501576	0.993097	0.501931	0.125444	0.117904	0.994069

UPDATED DATA PREPARATION

- Update to only look at admissions that lasted longer than 1 day
- Binning of (previously used) ICD-9 billing/diagnostic codes to diagnostic code category
- Cost Function

COST FUNCTION

- False Positive
  - Treat someone as if they are going to require readmission – by keeping him/her an additional day in the hospital.
  - For this project, the cost of a false positive was the cost of keeping a patient an additional day in the hospital. Average cost for non-profit facility is $2653. Cost for for-profit is $2093. Average for the two is $2373 x 75% (estimated cost to the hospital) = ***$1780 for an additional hospital day as the cost to the hospital.*** (Elflein, 2020)
- False Negative
  - Thought someone would not be readmitted and did not take additional action to prevent, but he/she actually was readmitted
  - **Average cost for readmission (any diagnosis) in 2016 was $14,400**. (Bailey, 2019)

COST- BASELINE

- Always predict negative (not readmitted)
  - Accuracy: 88.3%
  - Cost per prediction: $1,691.99
- Always predict positive (readmitted)
  - Accuracy: 11.7%
  - Cost per prediction: **$1,570.85**
- Random guess
  - Accuracy: 50%
  - Cost per prediction: $1,631.42

UPDATED RESULTS

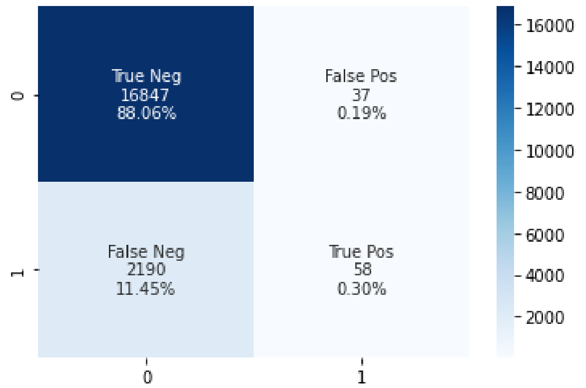
	Model	Train AUC	Train Accuracy	Train Precision	Train Recall	Train Cost	Validation AUC	Validation Accuracy	Validation Precision	Validation Recall	Validation Cost
5	RF	0.846940	0.846940	0.866088	0.820789	1403.270941	0.626888	0.635933	0.184783	0.615065	1218.837550
7	Ensemble	0.663879	0.663879	0.662713	0.667463	2696.602947	0.620703	0.610496	0.176957	0.634045	1235.971845
2	SGD	0.621266	0.621266	0.636039	0.566972	3406.552502	0.611095	0.649801	0.180723	0.560498	1275.065858
4	DTC	0.666799	0.666799	0.692863	0.599230	3121.955396	0.603574	0.656492	0.178593	0.534401	1301.855182
1	LogReg	0.595911	0.595911	0.561726	0.872826	1521.744325	0.580844	0.368179	0.140911	0.858837	1333.963342
0	KNN	0.607593	0.607593	0.622118	0.548122	3549.836718	0.588592	0.627779	0.165722	0.537367	1348.566451
6	GBC	0.792646	0.792646	0.797946	0.783751	1733.618744	0.585735	0.595512	0.159669	0.572954	1353.231584
3	NB	0.506505	0.506505	0.503322	0.985663	968.887561	0.502889	0.136665	0.118113	0.981613	1564.001673

Random Forest Classifier cost per prediction was $352.01 less than the best baseline predictor of always predicting positive.

RANDOM FOREST USING H20

- Skipped one-hot encoding of categorical variables
- Gain is marginal for splits on dummy variables
- One-hot encoding creates sparse decision trees
- H2O allows for true categorical variables
- Did not out-perform random forest in update results

```
Validation:
AUC:  0.5118046439552787
Accuracy:  0.8835981601505332
Precision:  0.6105263157894737
Recall:  0.025800711743772242
Confusion Matrix:
Cost:  1651.780263432992
```



COST SENSITIVE RANDOM FOREST USING COSTCLA

- Cost Matrix as input to algorithm
- Designed for observation dependent cost matrix
- Created loop to test different costs for FP and FN
- Did not out-perform random forest in update results

	FP	FN	Cost
11	6	6	1262.185518
33	16	16	1263.005087
89	41	46	1269.196460
0	1	1	1271.999442
78	36	41	1277.885567
...	...	...	...
80	41	1	1691.992473
38	16	41	1692.116524
48	21	41	1692.951425
58	26	41	1693.695728
36	16	31	1695.174577

CONCLUSION

Top Models

1. Random Forest Classifier cost per prediction was $1218.84, $352.01 less than the best baseline predictor of always predicting positive
2. Heterogeneous Ensemble of all updated models except Naïve Bayes and Random Forest was second best with $1235.97 per prediction
3. Cost Sensitive Random Forest using CostCla cost per prediction was $1262.19

Outstanding Questions

- Implications of predicting a readmission - keep one more day?
- Effect of keeping patient extra day on readmission rate

REFERENCES

- Bailey, M., Weiss, A., Barrett, M., & Jiang, J. (2019, February). Characteristics of 30-Day All-Cause Hospital Readmissions, 2010-2016. Retrieved from https://www.hcup-us.ahrq.gov/reports/statbriefs/sb248-Hospital-Readmissions-2010-2016.jsp#:~:text=The%20highest%20average%20readmission%20cost%20was%20for%20congenital,for%20pregnancy%2Fchildbirth%20%28%247%2C000%29%2C%20followed%20by%20mental%2Fbehavioral%20disorders%20%28%248%2C200%29
- Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore, "Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records," BioMed Research International, vol. 2014, Article ID 781670, 11 pages, 2014.
- Elflein, J. (2020, May 12). Inpatient day hospital costs in the U.S. by type 2018. Retrieved from https://www.statista.com/statistics/630443/inpatient-day-hospital-costs-in-us-by-nonprofit-or-profit/
- Long, A. (2020, January 30). Using Machine Learning to Predict Hospital Readmission for Patients with Diabetes with Scikit-Learn. Retrieved from https://towardsdatascience.com/predicting-hospital-readmission-for-patients-with-diabetes-using-scikit-learn-a2e359b15f0