University of Missouri System
COLUMBIA | KANSAS CITY | ROLLA | ST. LOUIS
UMKC S&T UMSL

Shalini Rallapalli Venkata <sr8k2@umsystem.edu>

IST 5520 Data Competition Final Result

Chen, Langtao <chenla@mst.edu> Fri, Apr 27, 2018 at 5:02 PM
To: "Bankar, Prasad (S&T-Student)" <phbdvz@mst.edu>, "Bollina, Avyakta (S&T-Student)" <abhbx@mst.edu>, "Chen, Zhe (S&T-Student)" <zc95c@mst.edu>, "Cui, Wenyuan (S&T-Student)" <wcz68@mst.edu>, "Huang, Junji (S&T-Student)" <jhyt7@mst.edu>, "Islam, Md Azharul (S&T-Student)" <mi6df@mst.edu>, "Kabir, Md Yasin (S&T-Student)" <mkdv6@mst.edu>, "Koli, Dhananjay Prakash (S&T-Student)" <dkgn9@mst.edu>, "Madra, Karan (S&T-Student)" <kmm7f@mst.edu>, "Mallapragada, Chandana (S&T-Student)" <cmdd8@mst.edu>, "Mao, Chenyi (S&T-Student)" <cm4z2@mst.edu>, "Mumudala, Kiran Kumar (S&T-Student)" <kmtq4@mst.edu>, "Rallapalli Venkata, Shalini (S&T-Student)" <sr8k2@mst.edu>, "Wang, Weiyu (S&T-Student)" <wwpmc@mst.edu>, "Wu, Hao (S&T-Student)" <hwxmc@mst.edu>, "Xu, Yongzhao (S&T-Student)" <yx26d@mst.edu>, "Yelamanchili, Tejaswini (S&T-Student)" <tyybf@mst.edu>, "Yerramareddy, Gautham (S&T-Student)" <gyfkr@mst.edu>, "Zhang, Miwen (S&T-Student)" <mz3z8@mst.edu>, "Zou, Cui" <tracyzou@mst.edu>, "Whitesides, Matthew B. (S&T-Student)" <mbwxd4@mst.edu>

Dear all,

Below is a summary of data competition II. Congrats to all students who have earned extra credit!

Student	F1 Score	Model
Miwen Zhang	0.9463	Random Forests + Hyperparameter Tuning Final model: RandomForestClassifier(n_estimators=150, max_features=5, criterion='entropy', random_state=123)
Junji Huang	0.9441	Gradient Boosting Classifier + Hyperparameter Tuning Final model: GradientBoostingClassifier(warm_start=True,learning_rate=0.30000000000000004,loss='exponential',n_estimators=300, random_state: 123)
Hao Wu	0.9432	Random Forests + Feature Selection + Hyperparameter Tuning + Class Weight Final model: RandomForestClassifier(n_estimators=200, max_features=4, criterion='gini', class_weight={1:1.5},random_state=123)
Yongzhao Xu	0.9388	Random Forests + Hyperparameter Tuning
Shalini Rallapalli Venkata	0.9383	Random Forests + Hyperparameter Tuning
Wenyuan Cui	0.9382	Random Forests + Hyperparameter Tuning
Cui Zou	0.9348	Random Forests + Hyperparameter Tuning + Class Weight
Md Yasin Kabir	0.9323	Neural Network (using keras package)
Zhe Chen	0.9161	Random Forests + Hyperparameter Tuning
Md Azharul Islam	0.9158	Neural Network + Hyperparameter Tuning
Avyakta Bollina	0.9004	Logistic regression + Feature Selection

A previous model with 0.99 F1 score has a technical issue and thus been removed. Among all 11 submissions, the best model has an F1 score as 0.9463 [the final model is RandomForestClassifier(n_estimators=150, max_features=5, criterion='entropy', random_state=123)].

To summarize, you can notice that all the top submissions are using ensemble methods (Random Forests and Gradient Boosting Machine) with hyperparameter tuning. You can also find that model tuning is very hard (and time consuming) but very important for business: even 1% improvement would make a huge difference for business success. Another option that has not been fully explored by all submissions is to extract more variables from the original dataset. With more information, you may build a better predictive model.

I think this is a useful exercise. If you are looking for a data scientist/engineer job, you will probably be asked by the recruiters to finish some data competitions similar to this one. I hope this exercise enhances your experience in building a predictive model with high performance.

The project M3 and M4 presentation slides are due this Sunday (April 29). Please submit your work and be prepared for the student presentation in the next week.

Best,

Langtao Chen

Ph.D., Assistant Professor

Department of Business and Information Technology

Missouri University of Science and Technology

106B Fulton Hall, 301 W. 14th Street, Rolla, MO 65409
chenla@mst.edu | https://bit.mst.edu/