no free hunch (http://blog.kaggle.com/) kaggle (http://www.kaggle.com)

▼ Archives

Tutorials and Winners' Interviews (http://blog.kaggle.com/category/dojo/)

Practice Fusion Diabetes Classification - Interviews with Winners

Posted on October (http://blog.kaggle.com/2012/10/) 3 (http://blog.kaggle.com/2012/10/03/) 2012 (http://blog.kaggle.com/2012/) by Margit Zwemer (http://blog.kaggle.com/author/margitzwemer/)

We check in with the 1st, 2nd, and 3rd place teams in the Practice Fusion Diabetes Classification Challenge (https://www.kaggle.com/c/pf2012-diabetes) (based on Shea Parkes' top voted (https://www.kaggle.com/c/pf2012/prospector) submission in the Prospect round). As an experiment, we've decided to group all the winners interviews together in one post to really highlight the diversity of backgrounds among successful data scientists.

What are your backgrounds prior to entering this competition?

1st place: Jose Antonio Guerrero (https://www.kaggle.com/users/5642/blind-ape) aka 'blind ape', Sevilla, Spain: My degrees are in mathematics, statistics and operations research. I'm worked in the public health sector for 25 years as researcher, IT technician and senior manager. A year ago, when I turned 50, decided it was a good age for return at my professional origin, so I went to Virgen del Rocio Universitary Hospital (http://huvr.es/), the flagship hospital in the region. I'm working with large size (8 figures) clinic records databases, grouping clinical cases and with quality and research issues

2nd: Matt Berseth (https://www.kaggle.com/users/26767/mtb) aka 'mtb', Jacksonville, FL, USA: I have a Bachelor's degree in computer science and a Master's degree in software engineering - both from North Dakota State University. I started my career as an intern with Microsoft and worked there full time for three years. Since leaving Microsoft, I have been working as a full stack developer with primarily Microsoft technologies for the last ten years. I have been fortunate to work in a variety of interesting areas including: automotive marketing, transportation management / logistics and healthcare IT.

3rd: Shashi Godbole (https://www.kaggle.com/users/4960/shashi-godbole) aka 'An apple a day', Mumbai, India: Data mining has been the focus of my work for the past six odd years. Earlier this year, I started a consulting firm with a friend from my alma mater. I had worked on a few other Kaggle competitions in the past (including the Heritage Health Prize (https://www.heritagehealthprize.com/c/hhp) which is still going on) but had done it mostly for fun. These earlier competitions allowed me to brush up my skills and learn the latest advances in machine learning. I also turned to R as my primary tool for analysis.

What made you decide to enter?

Jose: My experience in health sector. I'm used to clinical databases.

Matt: I studied machine learning as an undergraduate, but that was over ten years ago. So last fall when Coursera launched their machine learning course I decided to take the opportunity to get back up to speed. I enjoyed the course and took Daphane Koller's graphical models course this spring as a follow-up. With all of that theory under my belt, I decided I should apply it to something tangible like one of the Kaggle competitions.

Shashi: Healthcare is one of the core focus areas of my team at the consulting firm we are building. The problem statement of this particular competition resonated strongly with the kind of problems we are looking to solve for healthcare providers, payers and other stakeholders.

What preprocessing and supervised learning methods did you use?

Jose: Main work was data cleaning and feature creation. Grouping diagnostics and actives principles was crucial. As an advance of my solution, I based it in a hard preprocessing and feature creation work:

- Translating each medication to its active principles, route of administration.
- Grouping principles active by chemical families / clinical indication. In some cases, as statins, adjusting dose equivalences. Choosing for each group between number of prescriptions, dose or binary flag for feature creation.
- Grouping the diagnoses in base CCS and my personal experience.
- And much more...

After, the methods used were the well known gbm and randomforest and later stacking in a generalized additive model.

Matt: I spent a fair amount of time generating features from the ICD9, NDC and lab data. I used wikipedia heavily to learn more about diabetes and create features from the diagnoses and treatments that are related to diabetes.

All of the models I selected for my final submissions were boosted trees. I used anywhere from 5 to 13 different models and blended/combined their predictions to create my submissions.

Shashi: The preprocessing was limited to missing value imputation for a few fields in the data tables. I used random forests, gradient boosting and neural networks to build several models which were finally stacked together to generate a final solution.

What was your most important insight into the data?

Jose: The great numbers of comorbidities and symptoms associated with diabetes.

Matt: The ICD9 data is rich. There is information in the hierarchy of the codes (i.e. what level a specific code belongs to), information regarding the health of the individuals family (i.e. any of the 'family history of' codes) and information regarding the individuals behavior (i.e. any of the 'history of non-compliance' codes). And of course the conditions that are associated with each of the codes.

For the first month or so of the competition I focused almost solely on feature generation. And a majority of these features were derived from the ICD9 codes.

Shashi: One big insight was that the formats for some key fields were different in train and test datasets. This was causing a much larger error on test data than that on training data. I could not think of any explanation for this for quite a while. Fixing the formatting discrepancy resolved this issue and made the train and test errors consistent.

Where you surprised by any of your insights or any key features?

Jose: I'm surprised by gender overall impact. In the data, diabetes was much more prevalent in male (15%) than female (11%) but when fitting the model, the gender influence fell to 0.17%. Probably other comorbidities associated to gender would explain this reduction.

Matt: I was most surprised by the area's that I did not find interesting features. I did not find the lab data very useful and I thought the drug information would be more useful than it was. That surprised me. I would be interested in seeing what features the other competitors found in this area.

Shashi: I created several new features by taking ratios of different pairs of features. I was surprised by the extent to which these features improved my model. I created these features towards the very end of the competition. It helped me jump up a couple of places on the leaderboard.

Which tools did you use?

Jose: R

Matt:

1. .Net / C# for writing the logic that generates the features
2. SQL Server for storing the data as well as basic analysis (just sql queries from management studio)
3. Python and scikit-learn for training, testing and evaluating the models
4. R for graphical analysis (ggplot2)

Shashi: I used only R to do all the data processing and the modeling. Excel was used just a little bit to do some quick plots on the data.

What have you taken away from this competition?

Jose: I learned a lot about active principles and their interactions with diabetes.

Matt: Trust your cross validation scores and use the public leaderboard as a measurement of competitiveness. When I selected my final models for submission, I picked the models with the lowest cross-validation scores, not the ones with the lowest public leaderboard scores. This worked well for me in this competition, using this formula, I ended up picking my five best models.

Shashi: The most important learning was that feature engineering is of utmost importance. Perhaps even more than any fine-tuning of modeling algorithms. Allocating a lot of time to just review the data and create useful features can result in significant performance improvement.

Margit Zwemer (http://blog.kaggle.com/author/margitzwemer/) Formerly Kaggle's Data Scientist/Community Manager/Evil-Genius-in-Residence. Intrigued by market dynamics and the search for patterns.

Follow @MPZwemer

Tutorials and Winners' Interviews (http://blog.kaggle.com/category/dojo/)

aka (http://blog.kaggle.com/tag/aka/)

creation (http://blog.kaggle.com/tag/creation/)

Diabetes Classification (http://blog.kaggle.com/tag/diabetes-classificatio

FL (http://blog.kaggle.com/tag/fl/)

Fusion (http://blog.kaggle.com/tag/fusion/)

information (http://blog.kaggle.com/tag/information/)

Jacksonville (http://blog.kaggle.com/tag/jacksonville/)

machine learning (http://blog.kaggle.com/tag/machine-learning/)

Microsoft (http://blog.kaggle.com/tag/microsoft/)

Practice Fusion Diabetes Classification (http://blog.kaggle.com/tag/practi

work (http://blog.kaggle.com/tag/work/)

1 Comment   No Free Hunch   1 Login

Recommend   Share   Sort by Best

Join the discussion…

Ashley · a year ago

This is scary, I would stay away from them. http://www.forbes.com/sites/ka...

Reply · Share ›

ALSO ON **NO FREE HUNCH**   WHAT'S THIS?

If you can’t beat them, invite them

2 comments • 4 months ago

3rd Place interview from the KDD Cup 2014

7 comments • a year ago

Don't Miss These Scripts: Otto Group Product Classification

1 comment • a month ago

Predicting March Madness: 1st Place Winner, Zach Bradshaw

1 comment • 3 months ago

Subscribe   Add Disqus to your site   Privacy

no free hunch (http://blog.kaggle.com/)
is kaggle (http://www.kaggle.com)'s blog covering the sport of data science.
follow us on twitter (http://www.twitter.com/kaggle) & facebook (http://www.facebook.com/kaggle).

enter your email to join ou