Link to course materials online (data, KNIME workflows, PDFs, etc.). Link shared in the zoom chat: bit.ly/datasciencebootcamp\_aug2020

# Questions:

## Q: drag and drop visualization tools

A: I’m not sure I can recommend one “go to” data visualization tool. R has fantastic visualization. Python as well. Plotly is one package I’ve seen many use. There are other code frameworks that are very flexible, such as D3. D3 is arguably the best visualization framework out there.

## Q: what is your book?

My book: (amazon link):

https://www.amazon.com/Applied-Predictive-Analytics-Dean-Abbott/dp/1118727967/ref=sr\_1\_2?dchild=1&keywords=applied+predictive+analytics&qid=1597236165&sr=8-2

## NOTES 8/11

## Q: Is competing on analytics a book ?

[**https://www.amazon.com/Competing-Analytics-New-Science-Winning/dp/1422103323**](https://www.amazon.com/Competing-Analytics-New-Science-Winning/dp/1422103323)

## Q: from industry standard, is there a CRISP-DM certificate worth pursuing?

There is no CRISP-DM as far as I am aware. But it’s definitely worth reading the document. It isn’t a narrative—it’s better used while you are completing a project, or to explain to others what analysts do.

ftp://[ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf](http://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf)

## Q: what is the difference between conditioned statistics and regular statics ( mean, median, max..etc)?

“conditioned” statistics are the same statistics (mean, median, max, etc.) but computed on subsets of the data rather than the whole population. For example, if you could condition statistics of an input variable like “age” on another variable like “survived” (the target) or sex (male/female) to see what the differences are.

## Q: What software made the Parallel Coodinates Plot?

In the notes it was IBM SPSS Modeler. But most packages can do this.

**From Justin to Everyone: (11:36 AM)**

**Are there any rules of thumb for correlation of categorical variables?**

-- YES. the Chi-square statistic. For example, in KNIME, you can use the crosstab node

A picture containing drawing

Description automatically generated

To calculate the chi-square statistic for pairs of categorical levels and overall. The figure below has the resulting plot. In the plot, rows are pclass (values 1, 2, 3) and columns are survived (values 0 and 1). Note that in each cell, there are 3 values: the actual count, the expected count (if the variable values are distributed “evenly” or “randomly”), and the chi-square statistic. The chi-square statistic is large if the difference between the actual and expected is large for the population size.

What we see is that we would expect far fewer 1st class passengers to survive and far more to die if the data were distributed randomly. This tells us there is a a relationship between 1st class and surviving (they are “correlated” – not the right term, but I mean this in a colloquial sense).

Notice that 2nd class passengers don’t have this shift in expected vs. actual—they are about what you would expect (158 actual vs. 171 expected) so the chi-square statistic is low.

Finally, the overall chi-square statistic (like a weighted mean of the cell values) is high (127.8592) which gives a very small p-value. There is a statistically significant relationship between these two categorical variables that isn’t random.

A screenshot of a cell phone

Description automatically generated

## Q: Is there an IBM SPSS modeler tutorial that you recommend?

Well, at the risk of being self-serving, I’d argue the best “hands-on” book for IBM SPSS Modeler is the book I co-authored, “The IBM SPSS Modeler Cookbook”, <https://www.amazon.com/SPSS-Modeler-Cookbook-Keith-McCormick/dp/1849685460/ref=sr_1_2?crid=2VN9M587UMQJB&dchild=1&keywords=ibm+spss+modeler+cookbook&qid=1597903040&sprefix=IBM+SPSS+MOdeler+%2Caps%2C198&sr=8-2>

We wrote it because there was no book in the market that explained how to use Modeler.

## Q: Does normalizing the data mean fitting the data into a bell curve?

Normalizing usually refers to scaling the data to values that are consistent for all variables. Sometimes to a range [0,1] (i.e., from 0 to 1). This shouldn’t be confused with Normal distributions as they aren’t describing the same thing.

## Notes 8/12

Books I’ve referenced in the bootcamp

-- the Sage publications books are the “Green” books I referenced. One of them, on missing values, can be found here: <https://us.sagepub.com/en-us/nam/missing-data/book9419>. The sage publications page for all the books in this statistics series can be found here: <https://us.sagepub.com/en-us/nam/series/Series486>

-- statistical rules of thumb: <https://www.amazon.com/Statistical-Rules-Thumb-Gerald-Belle/dp/0470144483/ref=sr_1_1?dchild=1&keywords=statistical+rules+of+thumb&qid=1597904025&s=books&sr=1-1>

-- another stats book I really like and is easy to read is The Cartoon Guide to Statistics (it’s actually very well done): <https://www.amazon.com/Cartoon-Guide-Statistics-Larry-Gonick/dp/0062731025/ref=sr_1_2?crid=1C56Z7FNWRVX7&dchild=1&keywords=cartoon+guide+to+statistics&qid=1597904064&s=books&sprefix=cartoon+guide%2Cstripbooks%2C212&sr=1-2>

## Notes 8/13

## Q: autoML depends on the ensembles?

No. autoML automates the building of models but can use any model or models in the process. (ensembles included, but not necessarily there)

## Q: what is the secret sauce for ensemble methods and are so special powers?

In general, it’s creating more model complexity so a wider variety of patterns can be found.

For some, like Random Forests, it’s the resampling (for diversity of data), overfitting (for accuracy), random variable selection (to break up the greedy search).

For others, like GBMs, it’s underfitting and piecing together lots of simple models to build the more complex ensemble model.

## Q: neural networks vs ensemble? which route would you go

Both. ☺ Neither wins all the time.

That stated, I tend to build GBMs more than Neural Networks these days in practice. They are tremendously accurate and require less data preparation. It allows me to focus more on the data interpretation and less on building the models.

## Q: would not the weighting of the records introduce bias?

Yes. I assume this is for Boosting. Weighting records (whether for Boosting or bootstrap sampling) for a single model would introduce bias. But in an ensemble, the bias is overcome by combining lots of models (each biased in a different way)

## Q: bias and variance are inversely related?

Essentially, yes. They aren’t mathematically “inverses” but they work operationally like inverses in that low bias usually produces high variance and vice versa.

## Q: low bias = overfitting; low variance = underfit

Ofteh, yes. Low bias models are more likely to overfit. You can get lower and lower bias by overfitting. The ideal models, however, have low bias and are not overfit (these are good models!)

## Q: which one is more severe?

They can both cause problems. Or they can both be good. Low bias and low variance is the idea. Low bias and high variance indicates overfit models (need to simplify the models). High bias and low variance indicates underfit model (need to make the models more complex!). High bias and high variance is just bad (like a random model)

## Q: what is weak learner? is the same lazy learner

Yes, these are the same. Weak learners are very simple models, or as I was implying, modeling algorithms that can create very simple models. Decision trees that go 1 level deep (a Decision Stump) or even 2-3 levels deep are considered weak.

## Q: building decision tree rules sometimes is time consuming. how to build ensembles

Not typically. Trees build fast.

## Q: how the neural net fair in your ensemble

I’ve had good success building ensembles of neural networks via bagging or just building multiple NNets. Just make sure you overfit them and that they don’t generate highly correlated predictions. Then I just averaged the predictions.

## Q: what is GBM stands for?

Gradient Boosting Machine

## 09:41:30 From Tim Faulkner : Dean: One thing that might help in the firehose is getting a list of recommended books for differing levels

## management

I mentioned above the “Competing on Analytics” book. I also like

* “Mining Your Own Business: A Primer for Executives on Understanding and Employing Data Mining and Predictive Analytics” (<https://www.amazon.com/Mining-Your-Own-Business-Understanding/dp/0996712100/ref=sr_1_6?dchild=1&keywords=Jeff+Deal&qid=1597905344&s=books&sr=1-6>) and
* Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners (<https://www.amazon.com/Data-Mining-Machine-Learning-Practitioners/dp/1118618041/ref=sr_1_2?dchild=1&keywords=dean+analytics&qid=1597905405&s=books&sr=1-2>) –NOTE: I wrote the Foreword for this Jared Dean book
* Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die (<https://www.amazon.com/Predictive-Analytics-Power-Predict-Click/dp/1119145678/ref=sr_1_1?dchild=1&keywords=Eric+siegel&qid=1597905718&s=books&sr=1-1>)

## engineers wanting to move into Data Eng role

Sorry. No good suggestions here. Probably focus on Python, but I’d recommend data engineers who are going to support data science personnel learn basic machine learning so they know why they are building what they are building.

## Data Eng wanting to grow into Data Sci

That’s what this course can do, or my book (link above).

Other good books in this space include

* **General data science**
  + Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking (<https://www.amazon.com/Data-Science-Business-Data-Analytic-Thinking/dp/1449361323/ref=sr_1_2?crid=1CZTA7Q5VBL4B&dchild=1&keywords=data+science+for+business&qid=1597905625&s=books&sprefix=data+science+for+b%2Cstripbooks%2C207&sr=1-2>)
* **R**
  + Applied Predictive Modeling (<https://www.amazon.com/Applied-Predictive-Modeling-Max-Kuhn/dp/1461468485/ref=sr_1_2?crid=3HXV9I26DM98Z&dchild=1&keywords=applied+predictive+modeling&qid=1597905788&s=books&sprefix=applied+predictive%2Cstripbooks%2C207&sr=1-2>)

## Q: what happen if you calculate the wrong number of clusters for K-means

If you build a model with the “wrong” number of clusters, the clusters won’t be distinct from one another and you may get muddled interpretation of your data. But there’s really no “wrong” answer, but instead better and worse models from the perspective of distinct clusters.

## Q: k-mean is very similar to KNN

Bingo. Nice inference. They both use Euclidean distance. But of course the final model is thought of very differently (k-means being unsupervised and k-nn supervised).

## Q: how to do K-mean model calibration? techniques? rebuilding the whole model?

Rebuild if you don’t like the model you have. They are fast to build so this usually isn’t a problem.

## Q: how to did we convert K-means to supervised learning problems

Not done typically, but if you wanted to, you could build k-means model on all the data and overlay the target variable on top of the clusters. That value (the average value of the target variable in each cluster) is the score given to all records that fall into the cluster. For the Titanic data, this means the target variable, survived, is converted to a double and we compute the mean survived value for all records in each cluster.

## Q: Are you familiar with MoJo and PoJo forms?

I am unfamiliar with these, but they appear to be H2O specific terms (<https://h2o-release.s3.amazonaws.com/h2o/rel-ueno/2/docs-website/h2o-docs/pojo-quick-start.html>) At first glance, it looks like H2O models are converted into Java. That’s a typical pattern. Some tools create C or C++ code. Some Java. Some SQL. They are all useful.

## Titanic Data Info

Kaggle site overview: <https://www.kaggle.com/c/titanic/data>

## KDD Cup 1998 Data info

Original solicitation: <https://kdd.ics.uci.edu/databases/kddcup98/kddcup98.html>

KDD Site Overview: <https://www.kdd.org/kdd-cup/view/kdd-cup-1998/Tasks>

Notes on this data set: <https://docs.google.com/document/d/1Qm4ucR4wTK6vb7s6Q5nwdJYnFTQPC7f1MxsKzFxjSEg/edit>