Winning on Kaggle: Becoming a better data scientist

Data Science Retreat
March 18th, 2019

About me

- Undergraduate in Computer Science and Biology (U. of British Columbia)
- Master's in Machine learning (U. College London)
- Master's in Management (INSEAD)
- Ph.D. in Data Science and Business Analytics (U. of Amsterdam)
- Serious competitor on **Kaggle** (and other platforms) from 2015-2018
- Worked as a commercial data scientist on a variety of projects since 2015

Qingchen

Researcher and lecturer at University of Amsterdam
Amsterdam, Netherlands
Joined 7 years ago · last seen in the past day
https://www.qingchenwang.info/

Followers 164

Competitions Grandmaster

Competitions Summary

Competitions Grandmaster

Current Rank 157 of 102,312

Highest Rank 14

9 10 6

Competitions: 34
Solo: 19 (56%)
Team: 15 (44%)

Goals for today

1. Demystify **Kaggle** competitions
   - Why should you compete?

2. Provide a **framework** for excellent performance (top 10%) with minimal effort (5-10 hours per competition)
   - How to compete efficiently?

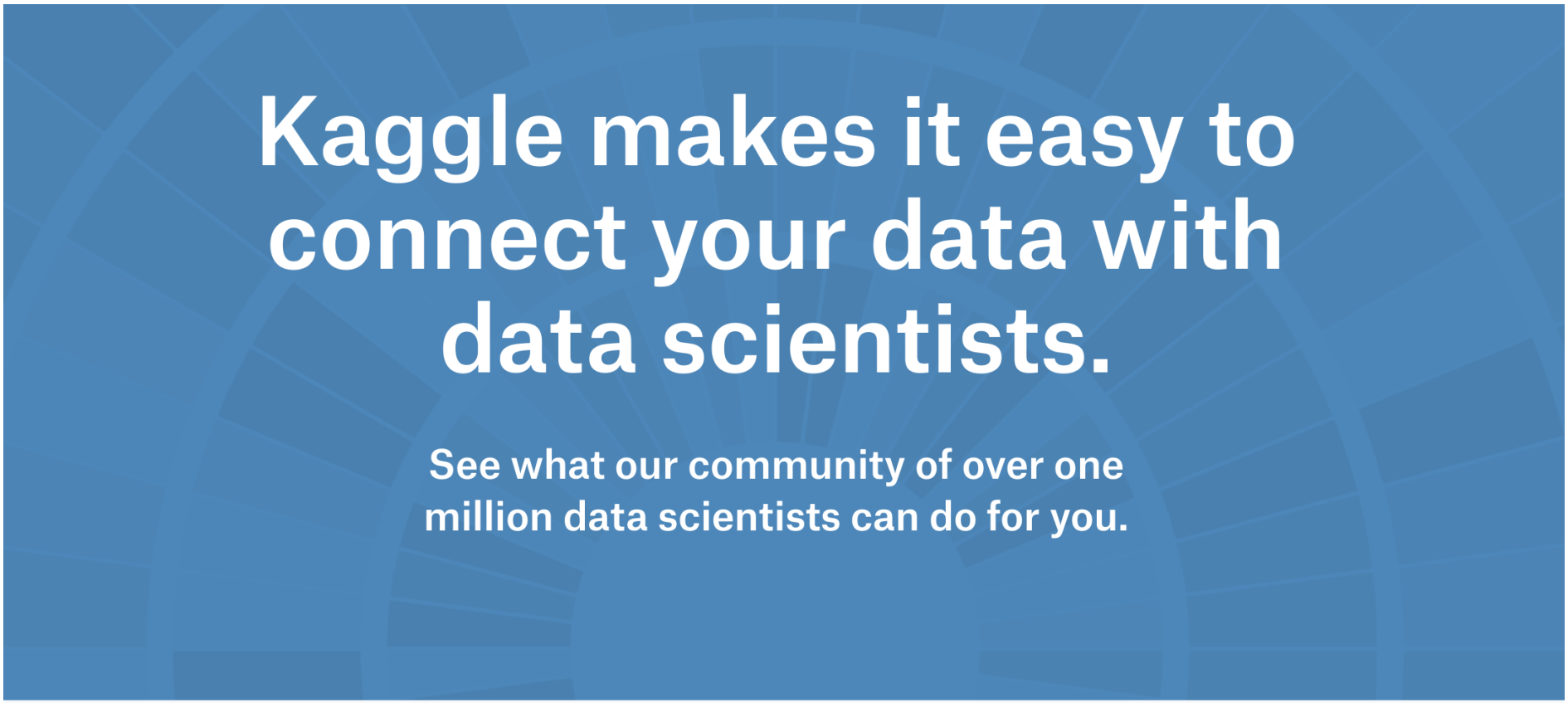
3. Provide some tips to **excel** in competitions
   - How to win?

4. Real life **experience** of a competition
   - I’ve set up a competition for you

Demystifying Kaggle competitions: Why you should compete

What do you think of Kaggle?

What do you think of Kaggle?



Kaggle is a platform for data science competitions.
We help you solve difficult problems, recruit strong teams, and amplify the power of your data science talent.

What do others think of Kaggle?

What do others think of Kaggle?

"Why Kaggle will NOT make you a great data scientist" (Pranay Dave 2018)

What do others think of Kaggle?

"Why Kaggle will NOT make you a great data scientist" (Pranay Dave 2018)

"Is Kaggle just for fun or is it something that I can write on my entry level data science resume?" (Quora 2016)

What do others think of Kaggle?

“Why Kaggle will NOT make you a great data scientist” (Pranay Dave 2018)

“Is Kaggle just for fun or is it something that I can write on my entry level data science resume?” (Quora 2016)

“5 Reasons Kaggle Projects Won't Help Your Data Science Resume” (Data Science Weekly)”

What do others think of Kaggle?

“Why Kaggle will NOT make you a great data scientist” (Pranay Dave 2018)

“Is Kaggle just for fun or is it something that I can write on my entry level data science resume?” (Quora 2016)

“5 Reasons Kaggle Projects Won't Help Your Data Science Resume” (Data Science Weekly)”

“No, Kaggle is unsuitable to study AI & ML. A reply to Ben Hamner” (Mostapha Benhenda 2017)

What do others think of Kaggle?

“Why Kaggle will NOT make you a great data scientist” (Pranay Dave 2018)

“Is Kaggle just for fun or is it something that I can write on my entry level data science resume?” (Quora 2016)

“5 Reasons Kaggle Projects Won't Help Your Data Science Resume” (Data Science Weekly)”

“No, Kaggle is unsuitable to study AI & ML. A reply to Ben Hamner” (Mostapha Benhenda 2017)

I used to agree with the above. I thought Kaggle was like World of Warcraft

What do others think of Kaggle?

“Why Kaggle will NOT make you a great data scientist” (Pranay Dave 2018)

“Is Kaggle just for fun or is it something that I can write on my entry level data science resume?” (Quora 2016)

“5 Reasons Kaggle Projects Won't Help Your Data Science Resume” (Data Science Weekly)”

“No, Kaggle is unsuitable to study AI & ML. A reply to Ben Hamner” (Mostapha Benhenda 2017)

I used to agree with the above. I thought Kaggle was like World of Warcraft

The truth is...

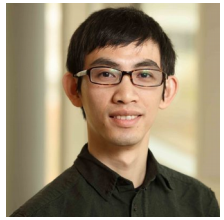
- Kaggle is a grind, but you learn and gain in real life from it!
- People are really nice!
- Kaggle will definitely make you a great data scientist!

Where are they now?









Giba – Lead Data Scientist at Ople.ai
Marios – Data Scientist at H2O.ai (London)
Xiaozhou (Little Boat) – Chief Data Scientist at Quartic.ai (Canada)
Mattias (Faron) – Data Scientist at H2O.ai (SV)
Abhishek Thakur – Chief Data Scientist at Boost.ai (Norway)
Carlos (NxGTR) – Senior Machine Learning Engineer at Instacart (SV)
Bojan – Data Scientist at H2O.ai (SV)
Jiwei – Data Scientist at Nvidia
Shize – Data/Applied Scientist at Microsoft
Vladamir – Computer Vision Engineer at Lyft
SRK – Data Scientist at H2O.ai (Chennai)
Jeong-Yoon – Senior Data Scientist Manager at Uber
Chenglong – Algorithm expert at Alibaba
Tom – Research engineer at DeepMind
Li – Research scientist at Google
Owen – Hedge fund something :) (Formerly Chief Product Officer at DataRobot

“Why Kaggle will NOT make you a great data scientist” (Pranay Dave 2018)

“Why Kaggle will NOT make you a great data scientist” (Pranay Dave 2018)

Claim: Data science is not only about prediction

- For example, many companies are interested to know what are the most common paths to customer churn?

“Why Kaggle will NOT make you a great data scientist” (Pranay Dave 2018)

Claim: Data science is not only about prediction

- For example, many companies are interested to know what are the most common paths to customer churn?

– Rebuttal: Prediction allows you to extract useful features to generalize a large dataset

- How can you summarize customer behavior when you have exponential number of paths? Even if you know the common paths to churn, how can you prevent it? Predict of course!

“Why Kaggle will NOT make you a great data scientist” (Pranay Dave 2018)

Claim: Data science is not only about prediction

- For example, many companies are interested to know what are the most common paths to customer churn?

– Rebuttal: Prediction allows you to extract useful features to generalize a large dataset

- How can you summarize customer behavior when you have exponential number of paths? Even if you know the common paths to churn, how can you prevent it? Predict of course!

Claim: You will not develop skills on graph algorithms

- Some real life problems are graph-based, and you can’t learn about algorithms like PageRank, Modularity, ShortestPath, EigenVectorCentrality.

“Why Kaggle will NOT make you a great data scientist” (Pranay Dave 2018)

Claim: Data science is not only about prediction

- For example, many companies are interested to know what are the most common paths to customer churn?

– Rebuttal: Prediction allows you to extract useful features to generalize a large dataset

- How can you summarize customer behavior when you have exponential number of paths? Even if you know the common paths to churn, how can you prevent it? Predict of course!

Claim: You will not develop skills on graph algorithms

- Some real life problems are graph-based, and you can’t learn about algorithms like PageRank, Modularity, ShortestPath, EigenVectorCentrality.

– Rebuttal: See Quora Question Pairs competition

- The truth is, graph algorithms are descriptive, they summarize a graph. What are you going to do with the summaries? Google use PageRank as a feature in predictions!

“Why Kaggle will NOT make you a great data scientist” (Pranay Dave 2018)

Claim: Data science is not only about prediction

- For example, many companies are interested to know what are the most common paths to customer churn?

– Rebuttal: Prediction allows you to extract useful features to generalize a large dataset

- How can you summarize customer behavior when you have exponential number of paths? Even if you know the common paths to churn, how can you prevent it? Predict of course!

Claim: You will not develop skills on graph algorithms

- Some real life problems are graph-based, and you can’t learn about algorithms like PageRank, Modularity, ShortestPath, EigenVectorCentrality.

– Rebuttal: See Quora Question Pairs competition

- The truth is, graph algorithms are descriptive, they summarize a graph. What are you going to do with the summaries? Google use PageRank as a feature in predictions!

Claim: You will not put effort in algorithm explicability

- People who got rejected for credit applications, have by law, a right to know why their application got rejected.

“Why Kaggle will NOT make you a great data scientist” (Pranay Dave 2018)

Claim: Data science is not only about prediction

- For example, many companies are interested to know what are the most common paths to customer churn?

– Rebuttal: Prediction allows you to extract useful features to generalize a large dataset

- How can you summarize customer behavior when you have exponential number of paths? Even if you know the common paths to churn, how can you prevent it? Predict of course!

Claim: You will not develop skills on graph algorithms

- Some real life problems are graph-based, and you can’t learn about algorithms like PageRank, Modularity, ShortestPath, EigenVectorCentrality.

– Rebuttal: See Quora Question Pairs competition

- The truth is, graph algorithms are descriptive, they summarize a graph. What are you going to do with the summaries? Google use PageRank as a feature in predictions!

Claim: You will not put effort in algorithm explicability

- People who got rejected for credit applications, have by law, a right to know why their application got rejected.

– Rebuttal: This has to do with the trade-off between model complexity and correctness

- Is it better to have a model that is easy to explain, but wrongly accepts/rejects applications 5% more often?

“Why Kaggle will NOT make you a great data scientist” (Pranay Dave 2018)

Claim: You will not get exposure to simulation and optimization

“Why Kaggle will NOT make you a great data scientist” (Pranay Dave 2018)

Claim: You will not get exposure to simulation and optimization

- Rebuttal: You learn this in OR courses, not in statistics, machine learning, or AI.

“Why Kaggle will NOT make you a great data scientist” (Pranay Dave 2018)

Claim: You will not get exposure to simulation and optimization

- Rebuttal: You learn this in OR courses, not in statistics, machine learning, or AI.

Claim: Missing link to Return on Investment (ROI) analysis

“Why Kaggle will NOT make you a great data scientist” (Pranay Dave 2018)

Claim: You will not get exposure to simulation and optimization

- Rebuttal: You learn this in OR courses, not in statistics, machine learning, or AI.

Claim: Missing link to Return on Investment (ROI) analysis

- Rebuttal: This is extremely important, but you won’t explicitly learn this in a PhD either.

“Why Kaggle will NOT make you a great data scientist” (Pranay Dave 2018)

Claim: You will not get exposure to simulation and optimization

- Rebuttal: You learn this in OR courses, not in statistics, machine learning, or AI.

Claim: Missing link to Return on Investment (ROI) analysis

- Rebuttal: This is extremely important, but you won’t explicitly learn this in a PhD either.

Claim: Deployment and operationalization cannot be experienced

“Why Kaggle will NOT make you a great data scientist” (Pranay Dave 2018)

Claim: You will not get exposure to simulation and optimization

- Rebuttal: You learn this in OR courses, not in statistics, machine learning, or AI.

Claim: Missing link to Return on Investment (ROI) analysis

- Rebuttal: This is extremely important, but you won’t explicitly learn this in a PhD either.

Claim: Deployment and operationalization cannot be experienced

- Rebuttal: How many companies have properly deployed data science commercially?
  - I have worked in data science for 4 years, nothing has been commercially deployed. This has nothing to do with data science and everything to do with a company’s strategy, management, and capability.
  - The 2nd largest publisher in the Netherlands spent five years building an excellent data lake but has no data science to show for.

“5 Reasons Kaggle Projects Won't Help Your Data Science Resume” (Data Science Weekly)”

Claim: Its hard to stand out

- Given the thousands of other people also doing them, it is becoming harder and harder for merely working through them to be enough to differentiate you.

“5 Reasons Kaggle Projects Won't Help Your Data Science Resume” (Data Science Weekly)”

Claim: Its hard to stand out

- Given the thousands of other people also doing them, it is becoming harder and harder for merely working through them to be enough to differentiate you.

– Rebuttal: That’s what most people think, so they spend very little effort. With the right skills, it is easy to get top 10% in any competition.

“5 Reasons Kaggle Projects Won't Help Your Data Science Resume” (Data Science Weekly)”

Claim: Its hard to stand out

- Given the thousands of other people also doing them, it is becoming harder and harder for merely working through them to be enough to differentiate you.

– Rebuttal: That’s what most people think, so they spend very little effort. With the right skills, it is easy to get top 10% in any competition.

Claim: You'll only demonstrate a partial Data Scientist skill-set

- Much of a data scientist's time is actually spent extracting, cleaning, and manipulating data, working on an independent project where you can showcase you can do this is more valuable.

“5 Reasons Kaggle Projects Won't Help Your Data Science Resume” (Data Science Weekly)”

Claim: Its hard to stand out

- Given the thousands of other people also doing them, it is becoming harder and harder for merely working through them to be enough to differentiate you.

– Rebuttal: That’s what most people think, so they spend very little effort. With the right skills, it is easy to get top 10% in any competition.

Claim: You'll only demonstrate a partial Data Scientist skill-set

- Much of a data scientist's time is actually spent extracting, cleaning, and manipulating data, working on an independent project where you can showcase you can do this is more valuable.

– Rebuttal: Many Kaggle competitions require cleaning and manipulating data.

- Data science is hard, don’t force yourself to be a unicorn. Work with what you have. If you don’t have an opportunity to extract data and develop an interesting independent project, be coming a Kaggle rockstar is the next best thing.

“5 Reasons Kaggle Projects Won't Help Your Data Science Resume” (Data Science Weekly)”

Claim: Its likely not something you're passionate about

“5 Reasons Kaggle Projects Won't Help Your Data Science Resume” (Data Science Weekly)”

Claim: Its likely not something you're passionate about

- Rebuttal: Do what you’re passionate about. Seriously try it, you’ll be hooked.

“5 Reasons Kaggle Projects Won't Help Your Data Science Resume” (Data Science Weekly)”

Claim: Its likely not something you're passionate about

- Rebuttal: Do what you’re passionate about. Seriously try it, you’ll be hooked.

Claim: It shows less initiative

- Hiring Manager's love to see candidates with enough drive to initiate a project in order to learn more or solve a problem they have personally faced.

“5 Reasons Kaggle Projects Won't Help Your Data Science Resume” (Data Science Weekly)”

Claim: Its likely not something you're passionate about

- Rebuttal: Do what you’re passionate about. Seriously try it, you’ll be hooked.

Claim: It shows less initiative

- Hiring Manager's love to see candidates with enough drive to initiate a project in order to learn more or solve a problem they have personally faced.

- Rebuttal: When you work hard on a Kaggle competition, you’ll know so much about the problem and be able to demonstrate initiative.

“5 Reasons Kaggle Projects Won't Help Your Data Science Resume” (Data Science Weekly)”

Claim: Its likely not something you're passionate about

- Rebuttal: Do what you’re passionate about. Seriously try it, you’ll be hooked.

Claim: It shows less initiative

- Hiring Manager's love to see candidates with enough drive to initiate a project in order to learn more or solve a problem they have personally faced.

- Rebuttal: When you work hard on a Kaggle competition, you’ll know so much about the problem and be able to demonstrate initiative.

Claim: It is likely irrelevant to the company

“5 Reasons Kaggle Projects Won't Help Your Data Science Resume” (Data Science Weekly)”

Claim: Its likely not something you're passionate about

- Rebuttal: Do what you’re passionate about. Seriously try it, you’ll be hooked.

Claim: It shows less initiative

- Hiring Manager's love to see candidates with enough drive to initiate a project in order to learn more or solve a problem they have personally faced.

- Rebuttal: When you work hard on a Kaggle competition, you’ll know so much about the problem and be able to demonstrate initiative.

Claim: It is likely irrelevant to the company

- Rebuttal: Most Kaggle competitions are highly relevant, data science problems are highly applicable to all industries.
  - Why would Facebook/Google/Microsoft/Amazon/Walmart etc host irrelevant competitions?

Other potential arguments

Claim: At the highest competitive level, Kaggle is neither for fun nor for resume...
Kaggle can actually become really stressful

Other potential arguments

Claim: At the highest competitive level, Kaggle is neither for fun nor for resume... Kaggle can actually become really stressful

- Rebuttal: Agreed, but everything is stressful when you compete to be the best in the world.
  - Fortunately, on Kaggle you get immediate feedback.
  - Submitting papers to peer review for my PhD is far more stressful. I spend 2 years on a paper and the reviewer rejects my paper because he or she doesn't agree with my premise.

Other potential arguments

Claim: At the highest competitive level, Kaggle is neither for fun nor for resume... Kaggle can actually become really stressful

- Rebuttal: Agreed, but everything is stressful when you compete to be the best in the world.
  - Fortunately, on Kaggle you get immediate feedback.
  - Submitting papers to peer review for my PhD is far more stressful. I spend 2 years on a paper and the reviewer rejects my paper because he or she doesn't agree with my premise.

Claim: Start with online courses

- You'll learn NLP, Neural Networks, etc.

Other potential arguments

Claim: At the highest competitive level, Kaggle is neither for fun nor for resume... Kaggle can actually become really stressful

- Rebuttal: Agreed, but everything is stressful when you compete to be the best in the world.
  - Fortunately, on Kaggle you get immediate feedback.
  - Submitting papers to peer review for my PhD is far more stressful. I spend 2 years on a paper and the reviewer rejects my paper because he or she doesn't agree with my premise.

Claim: Start with online courses

  - You'll learn NLP, Neural Networks, etc.
- Rebuttal: What do you think? I personally learn better from experience.
  - Many top Kagglers never received formal training in data science (Anokas, Giba, etc)

Other potential arguments

Claim: At the highest competitive level, Kaggle is neither for fun nor for resume... Kaggle can actually become really stressful

- Rebuttal: Agreed, but everything is stressful when you compete to be the best in the world.
  - Fortunately, on Kaggle you get immediate feedback.
  - Submitting papers to peer review for my PhD is far more stressful. I spend 2 years on a paper and the reviewer rejects my paper because he or she doesn't agree with my premise.

Claim: Start with online courses

- You'll learn NLP, Neural Networks, etc.

- Rebuttal: What do you think? I personally learn better from experience.
  - Many top Kagglers never received formal training in data science (Anokas, Giba, etc)

Claim: Do not waste your time on Kaggle

- There are so many original problems out there, with nobody to work on. Be the first at your own competition.

Other potential arguments

Claim: At the highest competitive level, Kaggle is neither for fun nor for resume... Kaggle can actually become really stressful

- Rebuttal: Agreed, but everything is stressful when you compete to be the best in the world.
  - Fortunately, on Kaggle you get immediate feedback.
  - Submitting papers to peer review for my PhD is far more stressful. I spend 2 years on a paper and the reviewer rejects my paper because he or she doesn't agree with my premise.

Claim: Start with online courses

- You'll learn NLP, Neural Networks, etc.

- Rebuttal: What do you think? I personally learn better from experience.
  - Many top Kagglers never received formal training in data science (Anokas, Giba, etc)

Claim: Do not waste your time on Kaggle

- There are so many original problems out there, with nobody to work on. Be the first at your own competition.

- Rebuttal: It's because so many people are working on the same problem that you can learn from each other. Your own competition does not allow you to learn from others.

Don’t try to be a unicorn!

1) Can you write code in a timely manner that works?
2) Can you write code to process raw data?
3) Can you train/validate/test models from known packages with reasonable performance?
4) Do you understand conceptually how some prediction or statistical models work?

Don’t try to be a unicorn!

1) Can you write code in a timely manner that works?
2) Can you write code to process raw data?
3) Can you train/validate/test models from known packages with reasonable performance?
4) Do you understand conceptually how some prediction or statistical models work?

Then you’re more useful than most data scientists I’ve met or worked with!

- Everything else is a matter of experience.

What I learned from Kaggle

What I learned from Kaggle

Creativity: Winning often requires doing something creative

What I learned from Kaggle

Creativity: Winning often requires doing something creative

Liberty Mutual: Treat a regression problem as a set of classification problems

What I learned from Kaggle

Creativity: Winning often requires doing something creative

Liberty Mutual: Treat a regression problem as a set of classification problems

BNP Paribas: One anonymous variable is actually user ID

What I learned from Kaggle

Creativity: Winning often requires doing something creative

Liberty Mutual: Treat a regression problem as a set of classification problems

BNP Paribas: One anonymous variable is actually user ID

Facebook Predict Check Ins: Use the future X variables to predict the past Y variable

What I learned from Kaggle

Creativity: Winning often requires doing something creative

Liberty Mutual: Treat a regression problem as a set of classification problems

BNP Paribas: One anonymous variable is actually user ID

Facebook Predict Check Ins: Use the future X variables to predict the past Y variable

Homesite Quote Conversion: Consider three-way featur interactions

What I learned from Kaggle

Data science “rigor”

- Learn to understand the prediction problem

What I learned from Kaggle

Data science “rigor”

- Learn to understand the prediction problem

How does the evaluation function affect performance of the model?

- AUC vs Logloss vs F1-score for classification problems.
- RMSE vs R-squared vs Gini coefficient.

What I learned from Kaggle

Data science "rigor"

- Learn to understand the prediction problem

How does the evaluation function affect performance of the model?

- AUC vs Logloss vs F1-score for classification problems.
- RMSE vs R-squared vs Gini coefficient.

How does time series affect the performance of the model?

- 10 months of training data, 2 months of test data, how should you validate your models?
- What about unobserved seasonality? Sales spikes in Nov. and Dec.

What I learned from Kaggle

Data science "rigor"

- Learn to understand the prediction problem

How does the evaluation function affect performance of the model?

- AUC vs Logloss vs F1-score for classification problems.
- RMSE vs R-squared vs Gini coefficient.

How does time series affect the performance of the model?

- 10 months of training data, 2 months of test data, how should you validate your models?
- What about unobserved seasonality? Sales spikes in Nov. and Dec.

Where does my model make mistakes?

What I learned from Kaggle

80/20 Rule of Data science

- If I describe to you a business problem, how fast can you design a data-driven solution?
- What should your target variable be?
- What data (features) do you need to predict the target?

What I learned from Kaggle

80/20 Rule of Data science

- If I describe to you a business problem, how fast can you design a data-driven solution?
- What should your target variable be?
- What data (features) do you need to predict the target?

Don’t waste precious time on difficult things

- In real life: You have 2 days to design and implement a good solution.
- In competitions: You have 30 days to beat everyone else.

What I learned from Kaggle

80/20 Rule of Data science

- If I describe to you a business problem, how fast can you design a data-driven solution?
- What should your target variable be?
- What data (features) do you need to predict the target?

Don’t waste precious time on difficult things

- In real life: You have 2 days to design and implement a good solution.
- In competitions: You have 30 days to beat everyone else.

What are the chances that this will improve performance?

- 100s of ideas to try, start with what you think is most likely to work.
- 100s of models to try, you should already know what works.

What I learned from Kaggle

The power of teamwork

- Wisdom of the crowd is powerful – especially when you take everyone's best ideas

What I learned from Kaggle

The power of teamwork

- Wisdom of the crowd is powerful – especially when you take everyone's best ideas

Less accomplished teammates are still valuable

- Every time I teamed up with another person my team's performance improved

What I learned from Kaggle

The power of teamwork

- Wisdom of the crowd is powerful – especially when you take everyone's best ideas

Less accomplished teammates are still valuable

- Every time I teamed up with another person my team's performance improved

Don't be afraid to ask for help

- Because less accomplished teammates are still valuable, you can add value.
- Home Depot Product Search – I was new to NLP problems and got stuck at 40th position, then joined a 6th ranked team, we finished 2nd.
- Quora Question Pairs – I was ranked 1st for a long time, then two others teamed up with me. I soon ran out of ideas, and my teammates were able to secure us a 5th place finish.

What I learned from Kaggle

Determination and grit is most important

- Many of the top Kagglers did not start strongly.
- It can be very daunting at first, but focus on one step at a time.
- After two or three competitions you'll all be experts.
- Some people took many years before finally getting their first gold medal.

What I learned from Kaggle

Determination and grit is most important

- Many of the top Kagglers did not start strongly.
- It can be very daunting at first, but focus on one step at a time.
- After two or three competitions you'll all be experts.
- Some people took many years before finally getting their first gold medal.

Data science is your career, think of everything you do now as an investment for the next 20-30 years.

Pick one or two things you're passionate about, and really do it!

What has Kaggle done for me?

What has Kaggle done for me?

Won some money

- $35000 total (between Kaggle and other platforms)

What has Kaggle done for me?

Won some money

- $35000 total (between Kaggle and other platforms)

Made some friends

- Marios, Giba, Carlos, Abhishek, Jiwei, Xiaozhou, Pawel, …, etc.

What has Kaggle done for me?

Won some money

- $35000 total (between Kaggle and other platforms)

Made some friends

- Marios, Giba, Carlos, Abhishek, Jiwei, Xiaozhou, Pawel, …, etc.

Get an industry job

- Had a nice offer as a Data Scientist this time last year – the interview was easy!

What has Kaggle done for me?

Won some money

- $35000 total (between Kaggle and other platforms)

Made some friends

- Marios, Giba, Carlos, Abhishek, Jiwei, Xiaozhou, Pawel, …, etc.

Get an industry job

- Had a nice offer as a Data Scientist this time last year – the interview was easy!

Help my friend get a job

- He did a Kaggle competition with me and we got a gold medal. His employer was very impressed during the interview

What has Kaggle done for me?

Get an academic job

- Many university departments were interested in my Kaggle experience

What has Kaggle done for me?

Get an academic job

- Many university departments were interested in my Kaggle experience

Get other opportunities

- Consulting projects, research projects, etc.

What has Kaggle done for me?

Get an academic job

- Many university departments were interested in my Kaggle experience

Get other opportunities

- Consulting projects, research projects, etc.

Brought me here today

- I met Jose at the KaggleDays event!

What has Kaggle done for me?

Get an academic job

- Many university departments were interested in my Kaggle experience

Get other opportunities

- Consulting projects, research projects, etc.

Brought me here today

- I met Jose at the KaggleDays event!

Most fun 2 years of my life!

Almost everything I learned in school was useless

Almost everything I learned in school was useless

What's your favorite prediction model?

- Mine used to be Bayesian probabilistic models (e.g., HMM, LDA).

What did I learn in school?

- Linear regression + regularization, SVM + Kernels, Perceptron, Adaboost, Probabilistic modeling.

What was my takeaway from school?

- Should try different methods for a given prediction problem because you don't know what might work.

Almost everything I learned in school was useless

What's your favorite prediction model?

- Mine used to be Bayesian probabilistic models (e.g., HMM, LDA).

What did I learn in school?

- Linear regression + regularization, SVM + Kernels, Perceptron, Adaboost, Probabilistic modeling.

What was my takeaway from school?

- Should try different methods for a given prediction problem because you don't know what might work.

All of this is wrong!

- Most prediction models are developed and published based on academic novelty and mathematical elegance.
- Academics traditionally did not care how well a model predicts as long as it beats a benchmark.

Everything I learned was useful

Everything I learned was useful

- Machine learning is for one purpose and one purpose only – prediction.
- To predict well is to minimize some error between the predicted and true values.
- A model needs to be evaluated by how well it predicts.

Everything I learned was useful

- Machine learning is for one purpose and one purpose only – prediction.
- To predict well is to minimize some error between the predicted and true values.
- A model needs to be evaluated by how well it predicts.
- Over-fitting happens – all the time!

Everything I learned was useful

- Machine learning is for one purpose and one purpose only – prediction.
- To predict well is to minimize some error between the predicted and true values.
- A model needs to be evaluated by how well it predicts.
- Over-fitting happens – all the time!
- Under-fitting happens – even more often!

How to achieve excellent performance with minimal effort

Getting started

1) Understanding the prediction problem.
2) Understanding the data.
3) Read the discussion forums.
4) Start with a public kernel.
5) Develop your own solution.

Getting started

1) Understanding the prediction problem.
2) Understanding the data.
3) Read the discussion forums.
4) Start with a public kernel.
5) Develop your own solution.

Follow the developments in discussions and public kernels and continue to work on your own solution.

Getting started

1) Understanding the prediction problem.
2) Understanding the data.
3) Read the discussion forums.
4) Start with a public kernel.
5) Develop your own solution.

Follow the developments in discussions and public kernels and continue to work on your own solution.

If your own solution can surpass the best public kernel, you will finish in the top 10%, most likely earn a silver medal.

Understanding the prediction problem

Using high-resolution lung scans to determine when lesions in the lungs are cancerous

Predicting how expensive a car insurance claim will be

Predicting faulty products from manufacturing lines

Predicting the relevance of search results

Predicting telecommunication network faults

Identifying who are likely to get cervical cancer

Predicting whether two questions are the same

Predicting stock returns

Forecasting sales at retail stores

Predicting prices of products on p2p markets

Answering exam questions

Understanding the data

Understanding the data

Data in a single file, M-by-N table, anonymized columns.

- Advantage: Very easy to get started.
- Disadvantage: Very difficult to be creative, lots of competition.

Understanding the data

Data in a single file, M-by-N table, anonymized columns.

- Advantage: Very easy to get started.
- Disadvantage: Very difficult to be creative, lots of competition.

Data in a single or set of files, column names provided

- Advantage: Fairly easy to get started, some room for creative.
- Disadvantage: Lots of competition.

Understanding the data

Data in a single file, M-by-N table, anonymized columns.

- Advantage: Very easy to get started.
- Disadvantage: Very difficult to be creative, lots of competition.

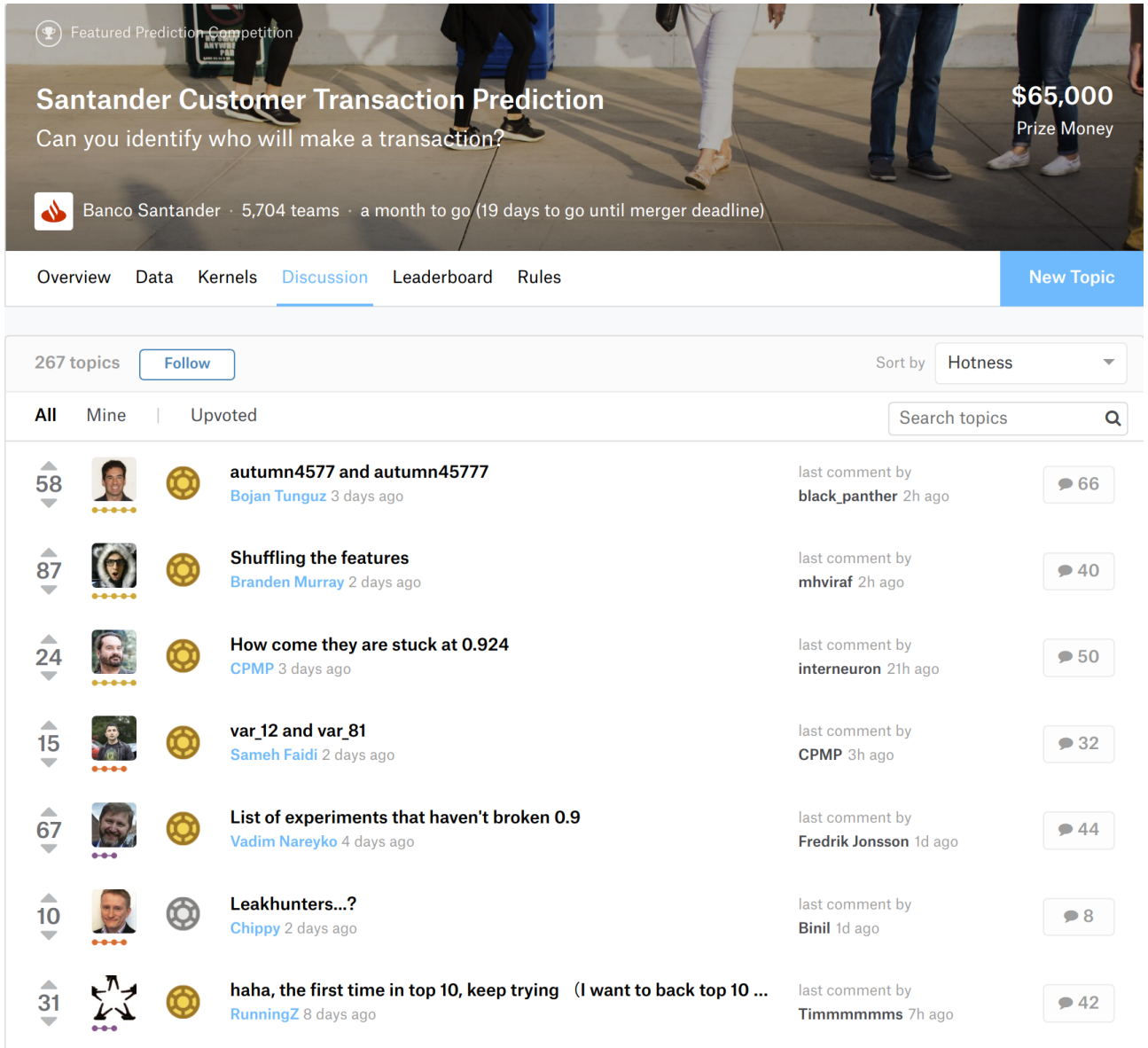
Data in a single or set of files, column names provided

- Advantage: Fairly easy to get started, some room for creative.
- Disadvantage: Lots of competition.

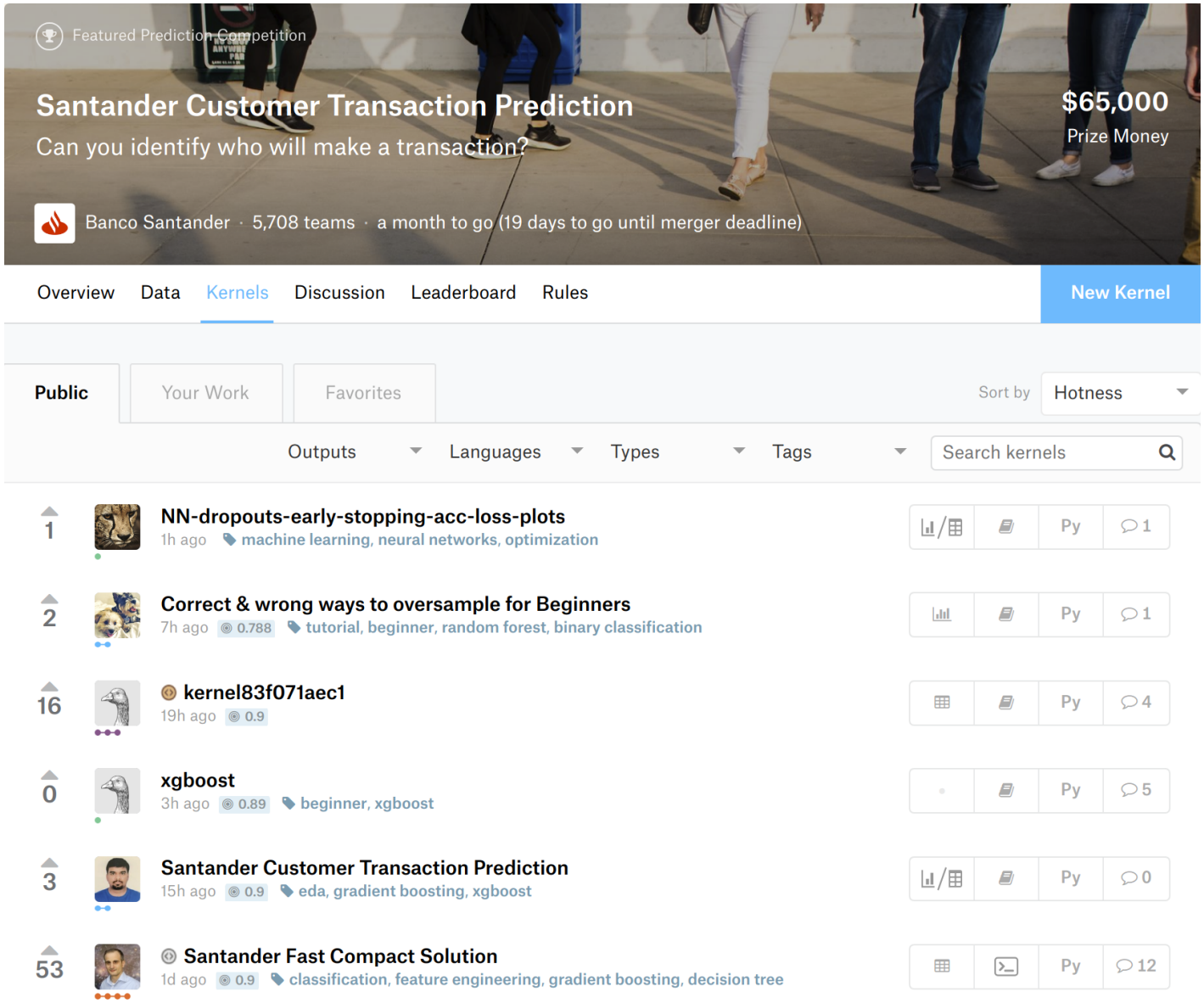
Data in unstructured format (e.g., images, unstructured text)

- Advantage: Freedom in how to put the data together. Weak competition
- Disadvantage: High entry barrier to get started.

Read the discussion forums



Start with the public kernel



Start with the public kernel

```
import numpy as np
import pandas as pd
import lightgbm as lgb
from sklearn.metrics import roc_auc_score
from sklearn.model_selection import StratifiedKFold
import warnings
warnings.filterwarnings('ignore')
train_df = pd.read_csv('../input/train.csv')
test_df = pd.read_csv('../input/test.csv')
features = [c for c in train_df.columns if c not in ['ID_code', 'target']]
target = train_df['target']
param = {
    'bagging_freq': 5,
    'bagging_fraction': 0.331,   'boost_from_average':'false',
    'boost': 'gbdt',             'feature_fraction': 0.0405,     'learning_rate': 0.0083,
    'max_depth': -1,             'metric':'auc',                 'min_data_in_leaf': 80,     'min_sum_hessian_in_lea
    'num_leaves': 13,            'num_threads': 8,               'tree_learner': 'serial',   'objective': 'binary',
}
folds = StratifiedKFold(n_splits=15, shuffle=False, random_state=2319)
oof = np.zeros(len(train_df))
predictions = np.zeros(len(test_df))
for fold_, (trn_idx, val_idx) in enumerate(folds.split(train_df.values, target.values)):
    print("Fold {}".format(fold_))
    trn_data = lgb.Dataset(train_df.iloc[trn_idx][features], label=target.iloc[trn_idx])
    val_data = lgb.Dataset(train_df.iloc[val_idx][features], label=target.iloc[val_idx])
    clf = lgb.train(param, trn_data, 1000000, valid_sets = [trn_data, val_data], verbose_eval=5000, early_stopping_
    oof[val_idx] = clf.predict(train_df.iloc[val_idx][features], num_iteration=clf.best_iteration)
    predictions += clf.predict(test_df[features], num_iteration=clf.best_iteration) / folds.n_splits
print("CV score: {:<8.5f}".format(roc_auc_score(target, oof)))
sub = pd.DataFrame({"ID_code": test_df.ID_code.values})
sub["target"] = predictions
sub.to_csv("submission.csv", index=False)
```

Liberty Mutual Group: Property Inspection Prediction

Competition background

A Fortune 100 company, Liberty Mutual Insurance has provided a wide range of insurance products and services designed to meet their customers' ever-changing needs for over 100 years.

To ensure that Liberty Mutual's portfolio of home insurance policies aligns with their business goals, many newly insured properties receive a home inspection. These inspections review the condition of key attributes of the property, including things like the foundation, roof, windows and siding. The results of an inspection help Liberty Mutual determine if the property is one they want to insure.

In this challenge, your task is to predict a transformed count of hazards or pre-existing damages using a dataset of property information. This will enable Liberty Mutual to more accurately identify high risk homes that require additional examination to confirm their insurability.

Liberty Mutual is interested in hiring predictive modelers like you to work on one of many growing analytics teams within our company. As a member of Liberty Mutual's advanced analytics community, you will have the opportunity to apply sophisticated, cutting-edge techniques, similar to those used in this competition, to large data sets in departments such as Actuarial, Product, Claims, Marketing, Distribution, Human Resources, and Finance. Click to view available positions.

This could have been you: Top 4%

70	▼ 56	STR	0.39302	174	4y
71	▼ 19	SummerRightnow	0.39302	22	4y
72	▲ 329	Sandro	0.39301	26	4y
73	▼ 25	Sad&PqiNNN	0.39301	122	4y
74	▼ 70	nrg	0.39300	228	4y
75	▼ 31	NxGTR	0.39300	154	4y
76	▼ 14	Daniel Chin	0.39297	81	4y
77	▼ 14	UNH Analytics	0.39294	56	4y
78	▼ 48	Hazardous ☠	0.39292	240	4y
79	▲ 384	Vincent Firmansyah	0.39291	34	4y
80	▲ 218	pi_informatics	0.39290	23	4y
81	▼ 8	Miner's Brain	0.39289	112	4y
82	▲ 153	YaTa	0.39289	6	4y
83	▲ 608	BlackMagic	0.39289	134	4y

General Takeaways

Best machine learning algorithms

- Logistic regression
  - Easy to understand and describe, performs reasonably well
- Gradient boosted decision trees
  - XGBoost, LightGBM
  - Great for everything, easy to use
- Feed-forward Neural Networks
  - Can give you great performance, but difficult to tune
- Convolutional Neural Networks
  - Best for images, can work well for text-based problems as well
- Recurrent Neural Networks
  - Seems to be good for text/speech problems
- Factorization Machines
  - LibFFM
  - Kills problems with high cardinality categorical variables

What to do for prediction problems

- Standard tabular problems
- Time series problems
- Text problems
- Image problems

Standard tabular data format

Standard tabular data format

- Think about how to handle categorical variables
  - Cardinality: high cardinality requires smart encoding

Standard tabular data format

- Think about how to handle categorical variables
  - Cardinality: high cardinality requires smart encoding
- Think about the distribution of numeric variables

Standard tabular data format

- Think about how to handle categorical variables
  - Cardinality: high cardinality requires smart encoding
- Think about the distribution of numeric variables
- Is it really a standard tabular data?
  - Maybe dates are involved? (Springleaf)
    - Differences in date values within row (start date, email date)
  - Maybe time series is involved? (Santander, Telstra)
    - Multiple rows for the same user implies sequence or time is involved

Time series data

Time series data

- How to validate model performance is key
  - Validation with time must reflect testing

Time series data

- How to validate model performance is key
  - Validation with time must reflect testing
- Look at the target variable across time periods
  - Trends, patterns

Time series data

- How to validate model performance is key
  - Validation with time must reflect testing
- Look at the target variable across time periods
  - Trends, patterns
- Combine predictions of individual entities and across entities
  - Rossmann

Text based problems

Text based problems

- Bags of words model is wonderful
  - One column for each word in a dictionary

Text based problems

- Bags of words model is wonderful
  - One column for each word in a dictionary
- Simple features are useful
  - Length of text, number of numerical values,
  - Don’t worry too much about spell check or text processing if data is sufficiently large

Text based problems

- Bags of words model is wonderful
  - One column for each word in a dictionary
- Simple features are useful
  - Length of text, number of numerical values,
  - Don’t worry too much about spell check or text processing if data is sufficiently large
- Embeddings will take it to the next level
  - LSA, LDA
  - Word2vec, GloVe, Fasttext

Text based problems

- Bags of words model is wonderful
  - One column for each word in a dictionary
- Simple features are useful
  - Length of text, number of numerical values,
  - Don’t worry too much about spell check or text processing if data is sufficiently large
- Embeddings will take it to the next level
  - LSA, LDA
  - Word2vec, GloVe, Fasttext
- Combine predictions of many different prediction algorithms
  - GBDT, FFM, CNN, RNN

Image problems

Image problems

- CNNs are exclusively used

Image problems

- CNNs are exclusively used
- Use pre-trained models (VGG, ResNet, Inception, etc)
  - Fine tune the model – retrain some of the layers

Image problems

- CNNs are exclusively used
- Use pre-trained models (VGG, ResNet, Inception, etc)
  - Fine tune the model – retrain some of the layers
- Augment the data
  - Do things like flip and rotate the images to enhance the training sample

Image problems

- CNNs are exclusively used
- Use pre-trained models (VGG, ResNet, Inception, etc)
  - Fine tune the model – retrain some of the layers
- Augment the data
  - Do things like flip and rotate the images to enhance the training sample
- Ensemble a lot!