Real World: Oscars
Predicting the Oscars

Poul Petersen
CIO, BigML
Where to Start?

Step 1

“Let’s predict customer churn!”

???

Step 2

Finish

“Here are the customers we predict will leave our service”
Where to Start?

“Let’s detect fraud!

???

“Here are the transactions we should stop immediately.
State the problem as an ML Task

• Remember: ML finds **patterns** in **data** enabling **predictions** about future events
• This means you need **data**
  • What **data** depends on what you want to **predict**
  • And the data you have or can collect
• Data needs to have **patterns** related to what you want to **predict**
  • Not magic: still can’t predict random events, lotteries, etc
• Your problem statement needs to be specific
  • Not “Let’s predict churn”
  • But “Let’s predict churn by looking at the profile data of all previous customers of our service who have/have not churned”
• This can be tricky…
Where to Start?

Step 1

“Let’s predict the Oscars!”

Step 2

???

Finish

“Here are the predicted winners”
Predicting the Oscars

BigML Scoresheet

2018

• 6 out of 6 right!
• 8 out of 8 actually, but probability of the predictions was “too low”
  • Adapted Screenplay
  • Original Screenplay

2019

• 4 our of 8 major awards correctly predicted
• Probabilities were lower this year
• This is still significantly better than guessing
Yay - Machine Learning!

There's no doubt that as I skim the #Oscars results, @bigmlcom and @pchh nailed the prediction game! The first with #MachineLearning and the latter with great assumption making about the voters.

What if there is a way to predict #Oscars before the awards? Wait! @bigmlcom just did that. With a 100% accuracy using #MachineLearning #AI, 3 days before. Impressive!!! #ML

#Oscar prediction from @bigmlcom is 100% accurate - power of #MachineLearning confirmed.

If that is true why can't we predict lottery numbers? We have winning history data available online.
Contrapositive Time

**Assertion:**

*We can predict the Oscars* $\Rightarrow$ *we can predict lottery numbers*

**Contrapositive:**

*We can’t predict lottery numbers* $\Rightarrow$ *we can’t predict the Oscars*

**Contradiction!**
Can’t Predict Lottery Numbers?

Nope Sorry! Two problems:

• The motion is chaotic (*that is extremely non-linear*)
• Even *small changes* in the initial conditions *greatly change* the outcome
• And you can’t measure the initial conditions with infinite precision  (thanks Heisenberg)

Question:
Why can we predict the Oscars?
Machine Learning’s ‘Amazing’ Ability to Predict Chaos

In new computer experiments, artificial-intelligence algorithms can tell the future of chaotic systems.

- ML to predict the propagation of a flame front
- This is also a chaotic system
- Succeeding in predicting out to 8 Lyapunov times
- Still a short amount of time
- A really short amount of time
- Lottery balls are allowed to “mix” for many, many Lyapunov times
- Does not contradict statement about predicting the lottery!
How an Oscar is Won

7,000+ members

Question:
Don’t we have the same problem as the lottery with predicting intention?
7,000 Chaotic Systems?

- Personal tastes
- Political considerations
- Values
- Cultural upbringing
- Critical education
- Pet peeves
- Corruption
Movie Watching Robots!

- Program robots to record audio and video
- Train them to react like a human to the movie
- Run a simulation and collect votes from the robots!

Please don’t try to solve everything with AI/ML
Ranking ML Applications

- **NO-BRAINERS**: Start here.
- **BRAINERS**: Development required.
- **NO-GO**: Not feasible.
- **POSTPONABLE**: Low ROI, no development required.

**Feasibility (inc data availability / dec complexity)**

**ROI (impact and cost)**

---

BigML, Inc

Real World Oscars
What if we Just Guess?

<table>
<thead>
<tr>
<th>Best Picture</th>
<th>Best Actor</th>
<th>Best Actress</th>
<th>Best Supporting Actor</th>
<th>Best Supporting Actress</th>
<th>Best Director</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Shape of Water</td>
<td>Gary Oldman</td>
<td>Frances McDormand</td>
<td>Sam Rockwell</td>
<td>Allison Janney</td>
<td>Guillermo del Toro</td>
</tr>
<tr>
<td>Dunkirk</td>
<td>Lady Bird</td>
<td>Margot Robbie</td>
<td>Woody Harrelson</td>
<td>Lesley Manville</td>
<td>Jordan Peele</td>
</tr>
<tr>
<td>The Post</td>
<td>daniel Day-Lewis</td>
<td>Sally Hawkins</td>
<td>Meryl Streep</td>
<td>Laura Linney</td>
<td>Christopher Nolan</td>
</tr>
<tr>
<td>Call Me By Your Name</td>
<td>The Post</td>
<td>Jennifer Lawrence</td>
<td>Christopher Plummer</td>
<td>Mary J. Blige</td>
<td>Paul Thomas Anderson</td>
</tr>
<tr>
<td>Get Out</td>
<td>Javon Carter</td>
<td>Viola Davis</td>
<td>Will Smith</td>
<td>Octavia Spencer</td>
<td></td>
</tr>
<tr>
<td>Three Billboards Out...</td>
<td>Denzel Washington</td>
<td>Viola Davis</td>
<td>Richard Jenkins</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dunkirk</td>
<td>Christopher Nolan</td>
<td>Frances McDormand</td>
<td>Denzel Washington</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Octopus</td>
<td>John Boyega</td>
<td>Frances McDormand</td>
<td>Denzel Washington</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1 of 9 1 of 5 1 of 5 1 of 5 1 of 5 1 of 5

28,125 COMBINATIONS

0.00003556 PROBABILITY
Consider Predicting Coin Tosses

After observing the first five flips as above, what is the probability that flip 6 is Heads? As stated, these events are independent, so the previous flips do not matter.
Consider Predicting Coin Tosses

What if this wasn’t a fair coin?
Idea: movies are not “equally” likely to be Best Picture…

Flip 1: Tails
Flip 2: Heads
Flip 3: Tails
Flip 4: Tails
Flip 5: Tails
Flip 6: ?Heads?

Prob <50%
Let’s Predict Best Picture

- These events are *not* independent
- Similar, but not identical, factors contribute to each win…
- We can expect a higher probability for Shape of Water to win
# The Features

## Data pulled from IMDB...

<table>
<thead>
<tr>
<th>MOVIES</th>
<th>AWARDS</th>
<th>OBJECTIVE</th>
</tr>
</thead>
<tbody>
<tr>
<td>• year</td>
<td>• Oscar_Best_Picture_nominated</td>
<td>• Oscar_Best_Picture_won</td>
</tr>
<tr>
<td>• movie</td>
<td>• Oscar_Best_Director_nominated</td>
<td>• Oscar_Best_Director_won</td>
</tr>
<tr>
<td>• movie_id</td>
<td>• Oscar_Best_Actor_nominated</td>
<td>• Oscar_Best_Actor_won</td>
</tr>
<tr>
<td>• certificate</td>
<td>• Oscar_Best_Actress_nominated</td>
<td>• Oscar_Best_Actress_won</td>
</tr>
<tr>
<td>• duration</td>
<td>• Oscar_Best_Supporting_Actor_nominated</td>
<td>• Oscar_Best_Supporting_Actor_won</td>
</tr>
<tr>
<td>• genre</td>
<td>• Oscar_Best_Supporting_ACTress_nominated</td>
<td>• Oscar_Best_Supporting_ACTress_won</td>
</tr>
<tr>
<td>• rate</td>
<td>• Oscar_Awards_screen_nominated</td>
<td>• Oscar_Awards_screen_won</td>
</tr>
<tr>
<td>• metascore</td>
<td>• Oscar_nominated</td>
<td>• Oscar_nominated</td>
</tr>
<tr>
<td>• synopsis</td>
<td>• Oscar_nominated_categories</td>
<td>• Oscar_nominated_categories</td>
</tr>
<tr>
<td>• votes</td>
<td>• Golden_Globes_won</td>
<td>• Golden_Globes_won</td>
</tr>
<tr>
<td>• gross</td>
<td>• Golden_Globes_won_categories</td>
<td>• Golden_Globes_won_categories</td>
</tr>
<tr>
<td>• release_date.year</td>
<td>• Golden_Globes_nominated</td>
<td>• Golden_Globes_nominated</td>
</tr>
<tr>
<td>• release_date.month</td>
<td></td>
<td>• Golden_Globes_nominated_categories</td>
</tr>
<tr>
<td>• release_date.day-of-month</td>
<td></td>
<td>• BAFTA_won</td>
</tr>
<tr>
<td>• release_date.day-of-week</td>
<td></td>
<td>• BAFTA_nominated</td>
</tr>
<tr>
<td>• awards Wins</td>
<td>• Screen_Actors_Guild_won</td>
<td>• Screen_Actors_Guild_won</td>
</tr>
<tr>
<td>• awards_nominations</td>
<td>• Screen_Actors_Guild_nominated</td>
<td>• Screen_Actors_Guild_nominated</td>
</tr>
<tr>
<td>• release_date.year</td>
<td>• Screen_Actors_Guild_nominated</td>
<td>• Screen_Actors_Guild_nominated</td>
</tr>
<tr>
<td>• release_date.month</td>
<td></td>
<td>• Screen_Actors_Guild_nominated</td>
</tr>
<tr>
<td>• release_date.day-of-month</td>
<td></td>
<td>• Critics_Choice_won</td>
</tr>
<tr>
<td>• release_date.day-of-week</td>
<td></td>
<td>• Critics_Choice_won</td>
</tr>
<tr>
<td>• awards Wins</td>
<td>• Critics_Choice_nominated</td>
<td>• Critics_Choice_nominated</td>
</tr>
<tr>
<td>• awards_nominations</td>
<td>• Critics_Choice_nominated</td>
<td>• Critics_Choice_nominated</td>
</tr>
<tr>
<td>• release_date.year</td>
<td>• Critics_Choice_nominated</td>
<td>• Critics_Choice_nominated</td>
</tr>
<tr>
<td>• release_date.month</td>
<td></td>
<td>• Critics_Choice_nominated</td>
</tr>
<tr>
<td>• release_date.day-of-month</td>
<td></td>
<td>• Critics_Choice_nominated</td>
</tr>
<tr>
<td>• release_date.day-of-week</td>
<td></td>
<td>• Critics_Choice_nominated</td>
</tr>
<tr>
<td>• awards Wins</td>
<td>• Directors_Guild_won</td>
<td>• Directors_Guild_won</td>
</tr>
<tr>
<td>• awards_nominations</td>
<td>• Directors_Guild_won</td>
<td>• Directors_Guild_won</td>
</tr>
<tr>
<td>• release_date.year</td>
<td>• Directors_Guild_nominated</td>
<td>• Directors_Guild_nominated</td>
</tr>
<tr>
<td>• release_date.month</td>
<td></td>
<td>• Directors_Guild_nominated</td>
</tr>
<tr>
<td>• release_date.day-of-month</td>
<td></td>
<td>• Directors_Guild_nominated</td>
</tr>
<tr>
<td>• release_date.day-of-week</td>
<td></td>
<td>• Directors_Guild_nominated</td>
</tr>
<tr>
<td>• awards Wins</td>
<td>• Producers_Guild_won</td>
<td>• Producers_Guild_won</td>
</tr>
<tr>
<td>• awards_nominations</td>
<td>• Producers_Guild_won</td>
<td>• Producers_Guild_won</td>
</tr>
<tr>
<td>• release_date.year</td>
<td>• Producers_Guild_nominated</td>
<td>• Producers_Guild_nominated</td>
</tr>
<tr>
<td>• release_date.month</td>
<td></td>
<td>• Producers_Guild_nominated</td>
</tr>
<tr>
<td>• release_date.day-of-month</td>
<td></td>
<td>• Producers_Guild_nominated</td>
</tr>
<tr>
<td>• release_date.day-of-week</td>
<td></td>
<td>• Producers_Guild_nominated</td>
</tr>
<tr>
<td>• awards Wins</td>
<td>• Art_Directors_Guild_won</td>
<td>• Art_Directors_Guild_won</td>
</tr>
<tr>
<td>• awards_nominations</td>
<td>• Art_Directors_Guild_won</td>
<td>• Art_Directors_Guild_won</td>
</tr>
<tr>
<td>• release_date.year</td>
<td>• Art_Directors_Guild_nominated</td>
<td>• Art_Directors_Guild_nominated</td>
</tr>
<tr>
<td>• release_date.month</td>
<td></td>
<td>• Art_Directors_Guild_nominated</td>
</tr>
<tr>
<td>• release_date.day-of-month</td>
<td></td>
<td>• Art_Directors_Guild_nominated</td>
</tr>
<tr>
<td>• release_date.day-of-week</td>
<td></td>
<td>• Art_Directors_Guild_nominated</td>
</tr>
<tr>
<td>• awards Wins</td>
<td>• Writers_Guild_won</td>
<td>• Writers_Guild_won</td>
</tr>
<tr>
<td>• awards_nominations</td>
<td>• Writers_Guild_won</td>
<td>• Writers_Guild_won</td>
</tr>
<tr>
<td>• release_date.year</td>
<td>• Writers_Guild_nominated</td>
<td>• Writers_Guild_nominated</td>
</tr>
<tr>
<td>• release_date.month</td>
<td></td>
<td>• Writers_Guild_nominated</td>
</tr>
<tr>
<td>• release_date.day-of-month</td>
<td></td>
<td>• Writers_Guild_nominated</td>
</tr>
<tr>
<td>• release_date.day-of-week</td>
<td></td>
<td>• Writers_Guild_nominated</td>
</tr>
<tr>
<td>• awards Wins</td>
<td>• Costume_Designers_Guild_won</td>
<td>• Costume_Designers_Guild_won</td>
</tr>
<tr>
<td>• awards_nominations</td>
<td>• Costume_Designers_Guild_won</td>
<td>• Costume_Designers_Guild_nominated</td>
</tr>
<tr>
<td>• release_date.year</td>
<td>• Costume_Designers_Guild_nominated</td>
<td>• Costume_Designers_Guild_nominated</td>
</tr>
<tr>
<td>• release_date.month</td>
<td></td>
<td>• Costume_Designers_Guild_nominated</td>
</tr>
<tr>
<td>• release_date.day-of-month</td>
<td></td>
<td>• Costume_Designers_Guild_nominated</td>
</tr>
<tr>
<td>• release_date.day-of-week</td>
<td></td>
<td>• Costume_Designers_Guild_nominated</td>
</tr>
</tbody>
</table>

**Engineered Features:**

- Award items field
- Nomination Counts
- Awards Counts
DATASET is publicly available:

https://bigml.com/user/academy_awards/gallery/dataset/5a94302592fb565ed400103b
Oscars Example

Tidbits and Lessons Learned....

• When specifying the problem, be as specific as possible
  • **Not:** “Let’s predict the Oscars”
  • **Instead:** “Let’s Predict the Oscars by correlating a series of award wins with the final Oscar win.”

• The statement of the problem will guide the data required

• Be aware of the cost of collecting the data versus the ROI:
## Effort of a ML Application

<table>
<thead>
<tr>
<th>Task</th>
<th>Effort</th>
</tr>
</thead>
<tbody>
<tr>
<td>State the problem as an ML task</td>
<td>~10%</td>
</tr>
<tr>
<td>Data wrangling</td>
<td>~80%</td>
</tr>
<tr>
<td>Data transformations</td>
<td></td>
</tr>
<tr>
<td>Feature engineering</td>
<td>~5%</td>
</tr>
<tr>
<td>Modeling and Evaluations</td>
<td>~5%</td>
</tr>
<tr>
<td>Predictions</td>
<td></td>
</tr>
<tr>
<td>Measure Results</td>
<td></td>
</tr>
</tbody>
</table>

This is only such low effort because of platforms like BigML® being innovating.
Oscars Example

Tidbits and Lessons Learned….

• When specifying the problem, be as specific as possible
  • Not: “Let’s predict the Oscars”
  • Instead: “Let’s Predict the Oscars by correlating a series of award wins with the final Oscar win.”
• The statement of the problem will guide the data required
• Be aware of the cost of collecting the data versus the ROI:
  • IMDB data is readily available
  • Start small and go straight to the desired result
• We’re done right?
  • Nope. You can’t escape Feature Engineering
  • Items: BAFTA_won_categories = list of nominations
  • Aggregations: Nomination and Award counts
• You can’t escape Feature Selection
  • Full user reviews costly to collect and not useful

Wait: How were you confident in the predictions?
Evaluating the Models

- Ultimately, we want to use all the history to predict the winner for the current year.
- In order to evaluate success, we use a model built from 2000-2012 data to predict the winners for 2013-2016.
- Built a separate Deepnet for each award category.
- Evaluation obtained a ROC AUC over 0.98 across all award categories.
Reality Check

*Three Important Concepts in Applying ML...*

- All Machine Learned models are **wrong**
- Real-world Machine Learning is **iterative**
- End-to-end Machine Learning is **compositional**
Tenets of Machine Learning

- All Machine Learned models are **wrong**, but some are useful
  - Better features always beat better algorithms
  - Good algorithms already exist and are good enough
  - Tools like OptiML exist which can help optimize performance
  - The data is **never** good enough
- Real-world Machine Learning is **iterative**
  - Automation is better than hand tuning - you need an API!
  - When data changes quickly, training speed is more important than accuracy
  - **Repeatability** is superior to a single strong result
- End-to-end Machine Learning is **compositional**
  - Problems are solved with workflows of algorithms
  - A ML solution is not real until it is in production
  - ML is here: Now we need 100,000x people applying ML
Your Turn!

• What are some problems you can solve with ML?
• Do you have the data
  • Where is it? Can you get it?
  • Does it need cleaning (hint: yes)
  • What ML tasks will be involved?
• Remember: go straight to the result
  • Prove it before you build it
  • Use Models and Logistic Regressions to start
  • Spend time on features and introspection