Business Analytics @ UF

DBA Research Symposium

Dr. Jim Hoover

jim.hoover@warrington.ufl.edu

352-294-0410

Objectives

- Background – From DBA to UF
- New Analytics Degree, Concentration, and Undergraduate updates
- A standard approach to Analytics projects that increases ROI
- Walking through a machine learning example – Recommender Engines

Videos and Spreadsheet

- The videos are hosted on this YouTube channel:
  - DBA Presentations: http://bit.ly/DBAvideos
    - Part 1 – Business Analytics at UF: https://youtu.be/v-t7ug-o3Ic
    - Part 2 –Marketing Analytics - Collaborative Filtering and Recommendation Engines: https://youtu.be/uAM6s-2qmuI
    - Part 3 – Walking through a recommendation engine example in Excel: https://youtu.be/DTvBVgvgAWQ
  - The spreadsheet example is hosted in this location:
    - https://github.com/Hoover-code/DBA/blob/master/CollaborativeFilteringExample.xlsx

https://warrington.ufl.edu/directory/person/7650/

Introduction: From DBA to UF Professor

Marketing Analytics Methodologies

- Standard Analytics Methodologies:
  - Permit teams across an organization to have a common approach to analytics projects
  - Help non-analytics professionals understand analytics projects
  - Help managers make decisions about investments in analytics projects, people, data and software
  - Facilitate analytics team members to make decisions and take actions during an analytics project
  - Describe in detail both large and small steps to consider taking in your analytics project
  - Identify responsibilities of team members who conduct the analytics project
- Major analytics methodologies used by firms and organizations:
  - CRoss Industry Reference Process for Data Mining (CRISP-DM) – Hyperlink and Summary
  - Sample, Explore, Modify, Model, and Assess (SEMMA) - Hyperlink
  - Foundation Methodology for Data Science (FMDS) – Hyperlink & Summary
  - Agile Analytics Project Model - Hyperlink
  - Predictive Analytics Lifecycle - Hyperlink
  - Team Data Science Process (TDSP) – Hyperlink & Summary
  - Firm-specific methodologies

The Cross Reference Industry Standard Process for Data Mining (CRISP-DM) – SPSS / IBM

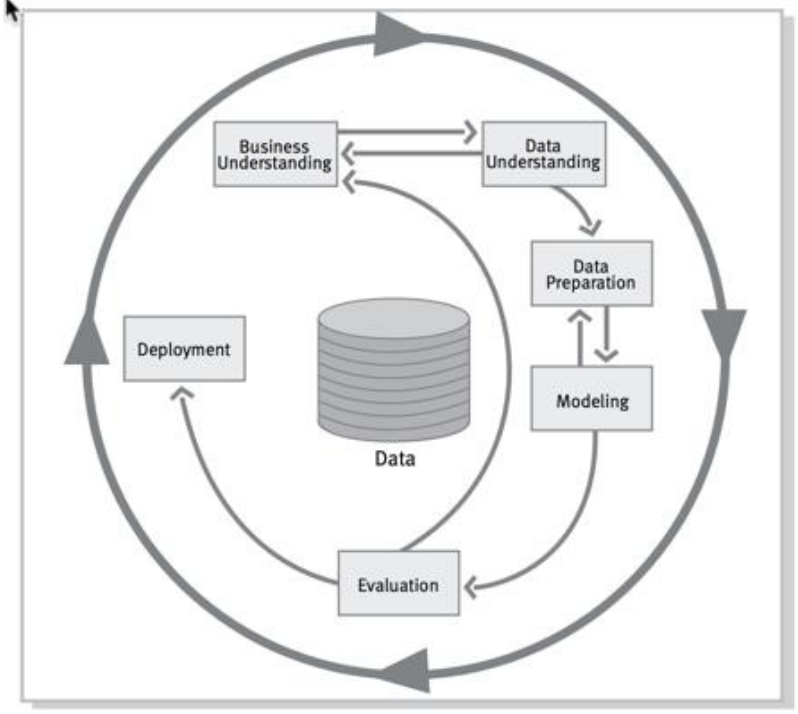


Figure 2: Phases of the CRISP-DM reference model

Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives *Background* *Business Objectives* *Business Success Criteria* **Assess Situation** *Inventory of Resources* *Requirements, Assumptions, and Constraints* *Risks and Contingencies* *Terminology* *Costs and Benefits* **Determine Data Mining Goals** *Data Mining Goals* *Data Mining Success Criteria* **Produce Project Plan** *Project Plan* *Initial Assessment of Tools and Techniques*	**Collect Initial Data** *Initial Data Collection Report* **Describe Data** *Data Description Report* **Explore Data** *Data Exploration Report* **Verify Data Quality** *Data Quality Report*	**Select Data** *Rationale for Inclusion/ Exclusion* **Clean Data** *Data Cleaning Report* **Construct Data** *Derived Attributes* *Generated Records* **Integrate Data** *Merged Data* **Format Data** *Reformatted Data* *Dataset* *Dataset Description*	**Select Modeling Techniques** *Modeling Technique* *Modeling Assumptions* **Generate Test Design** *Test Design* **Build Model** *Parameter Settings* *Models* *Model Descriptions* **Assess Model** *Model Assessment* *Revised Parameter Settings*	**Evaluate Results** *Assessment of Data Mining Results w.r.t. Business Success Criteria* *Approved Models* **Review Process** *Review of Process* **Determine Next Steps** *List of Possible Actions* *Decision*	**Plan Deployment** *Deployment Plan* **Plan Monitoring and Maintenance** *Monitoring and Maintenance Plan* **Produce Final Report** *Final Report* *Final Presentation* **Review Project** *Experience Documentation*

Figure 3: Generic tasks (bold) and outputs (italic) of the CRISP-DM reference model

Image captured from: https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.crispdm.help/crisp_overview.htm

The Cross Reference Industry Standard Process for Data Mining (CRISP-DM) – SPSS / IBM

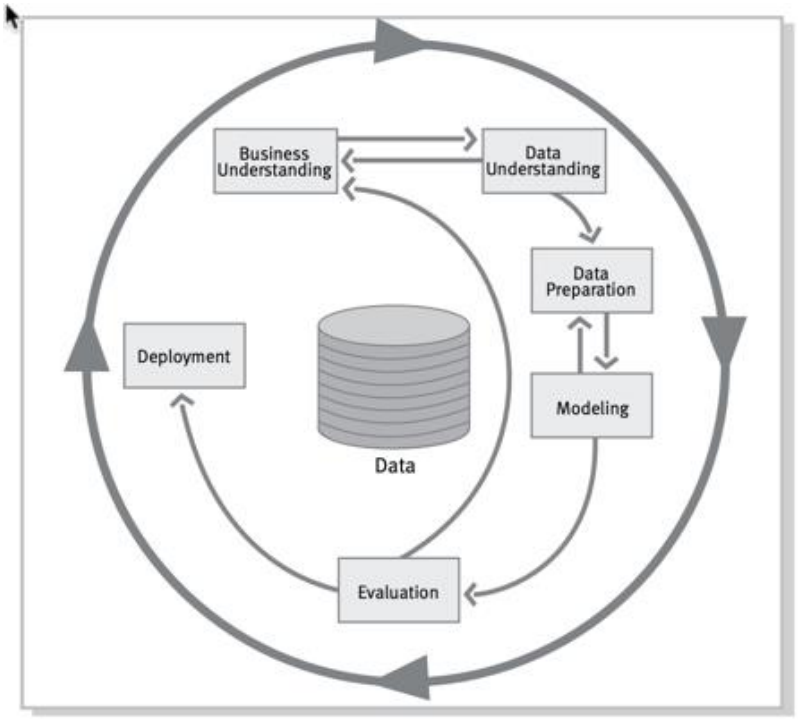


Figure 2: Phases of the CRISP-DM reference model

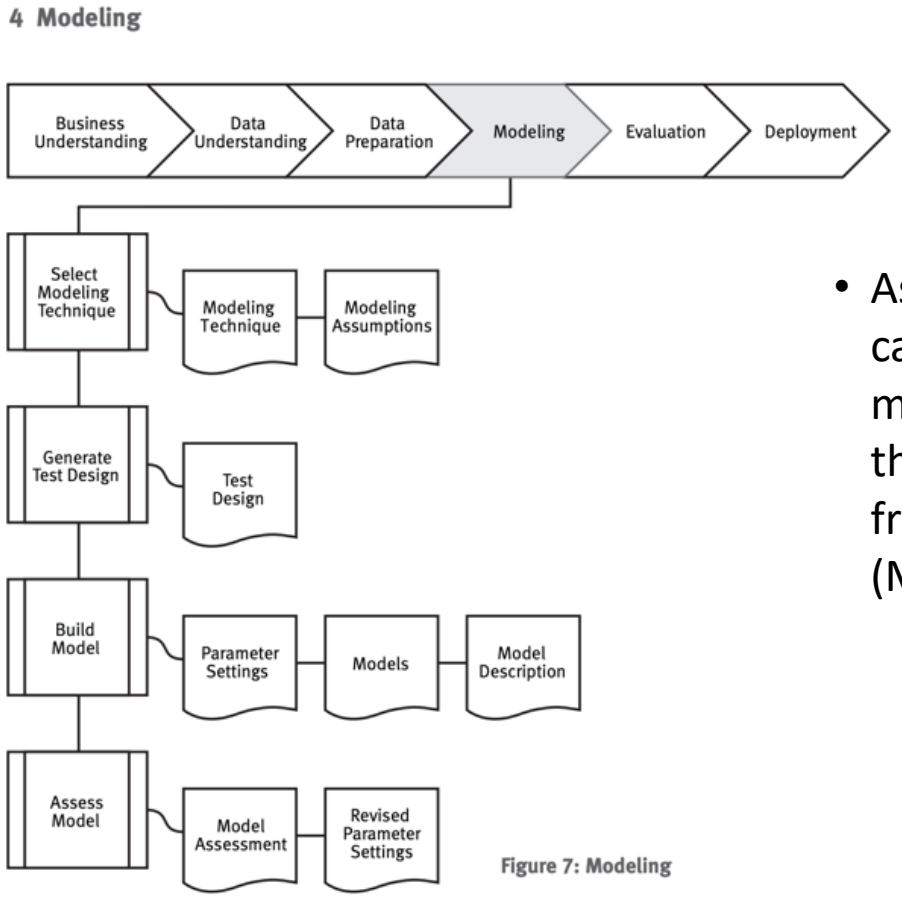


Figure 7: Modeling

- As you can see in this screen capture from the CRISP-DM model, the graphic shows the sub-steps and outputs from the major step (Modeling)

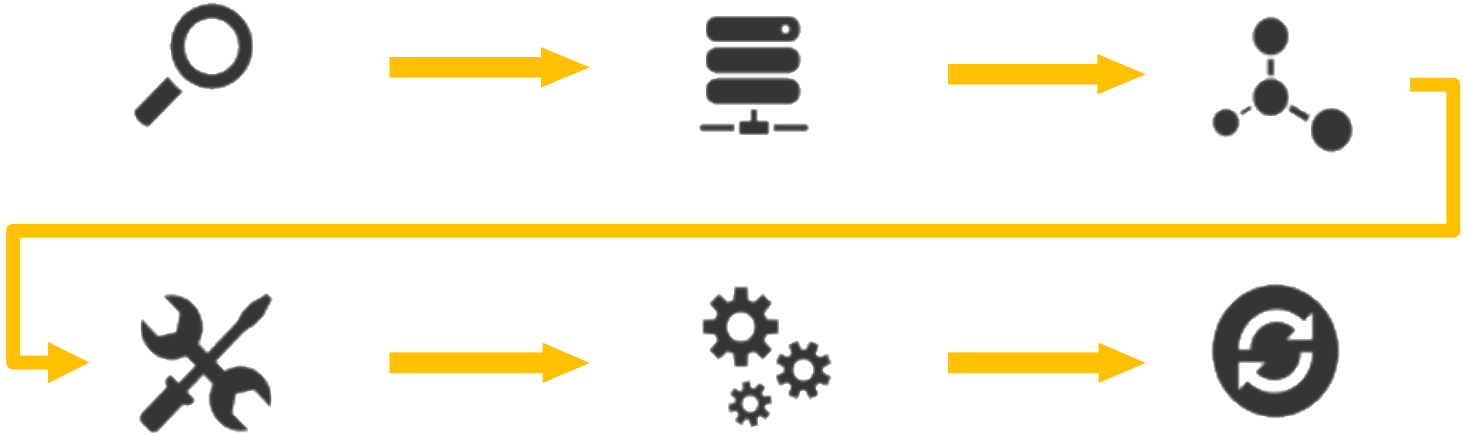
Image captured from: https://www.the-modeling-agency.com/crisp-dm.pdf

A Firm-Specific Analytics Methodology

1 **Understand the Business Problem:** Evaluate the key question the decision-makers face; assess how they will make a decision; assess the degree of precision required. How much of the action will be determined by a human?

2 **Evaluate Data Sources:** What are the most actionable and relevant data sets can be used to address the client problem? Includes internal data as well as external sources of data. Data importance to the question, access, missing and bad data, and data ingestion are considered.

3 **Assess the Best Analytical Techniques to Address the Problem:** What are the best analytical methods for achieving the desired results? Depending on: the analytic question; the data; and the level of precision needed by the decision-maker, choose the right modelling technique.



4 **Select the Right Technology:** Determine the best fitting set of tools to perform the analysis. Consider tools that are already employed, including: COTS, open source, and custom-built that best achieves desired results and reduces costs.

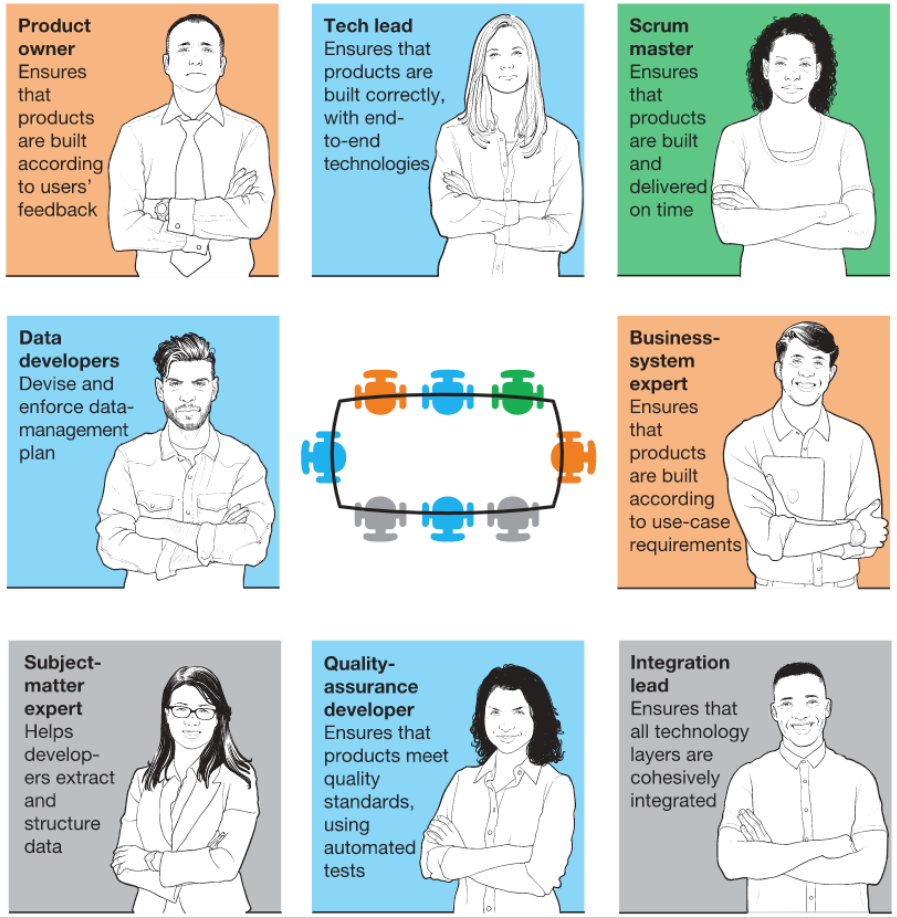
5 **Produce Analytic Outcomes:** Use the technologies, techniques, and data to produce analytic solutions that solve the client problem. Create analytic results that the client can act on. Validate results match the expected outcomes, can be understood and can be effectively implemented.

6 **Integrate with Business Processes and Systems:** Ensure that analytics outcomes are incorporated back into the processes and business systems in order to drive business value and achieve desired results. In the case of ML / AI, institute a continuous learning approach.

- Methodology used by a leading consulting firm
- Includes a focus on integration back into business systems and processes for implementation
- Includes most of the activities of CRISP-DM

The Analytics / Data Science Team

Agile Team from McKinsey



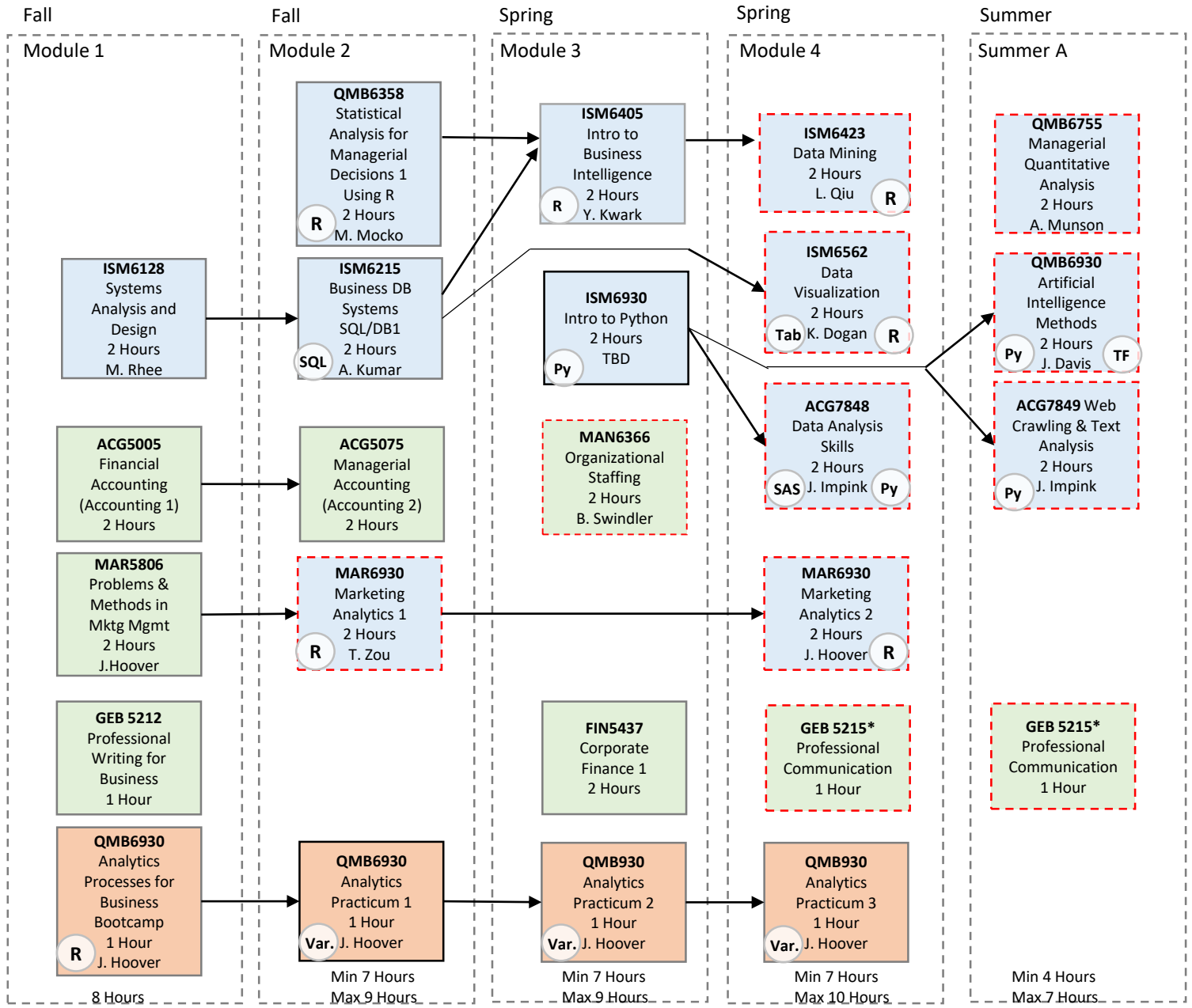
Typical Analytics Project Team

Data Scientist / Project Lead	• Combines domain expertise and analytics methodologies to drive informed analytics strategies for solution development • Validates the model meets requirements • Directs the data engineer in data munging efforts • Manages project team, timeline, scope, and delivery
Functionalist / Subject Matter Expert	• Maintains deep functional knowledge and/or industry experience in the project area • Brings project-specific knowledge of current of current practices, processes, and problems • Understands firm databases and data sources • Serves as the primary business POC for business understanding • Briefs the business on model outcomes
Statistician / Data Analyst	• Contributes statistical expertise and analysis skills required for data investigation and modeling • Performs model development
Data Engineer / DBA	• Develops data and solution architecture • Completes data wrangling and data munging efforts
Technical or Business Analyst	• Assists with data preparation • Helps build the proof-of concept technical solution • Documents artifacts for the project • Helps prepare the business briefing of model outcomes

Image captured from: https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/using-agile-to-accelerate-your-data-transformation

MSBA Track Curriculum

V3.1.4
3/9/2020



Legend

- Analytics Course – Technical
- Elective courses to achieve 36-hour degree requirements
- Business Course – Functional
- Elective Business Course – Functional
- Analytics Practicum Courses

R – R language
Py – Python language
SAS – SAS program
Tab – Tableau
SQL – Structured Query language
TF – TensorFlow
Var. – Various (project dependent)

36 Total Hours Required for Degree

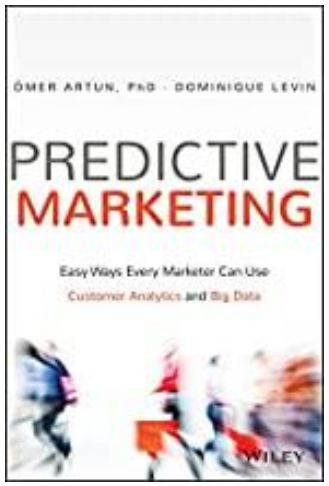
STEM degree CIP code 11.0802

* Note: GEB 5215 is required either Module 4 or Summer A

Marketing Analytics

- Some recommended books on Marketing Analytics:

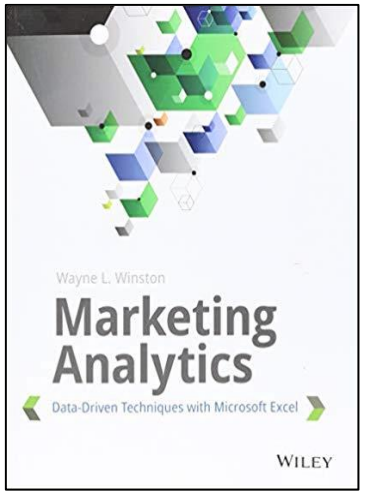
Artun, O. and Levin, D. (2015). *Predictive Marketing: Easy Ways Every Marketer Can Use Customer Analytics and Big Data*. Hoboken, New Jersey: John Wiley & Sons



Focuses on the benefits of predictive marketing efforts without the modeling details.

This book is available at Amazon.com. The approximate cost should be approximately $25 - $30.

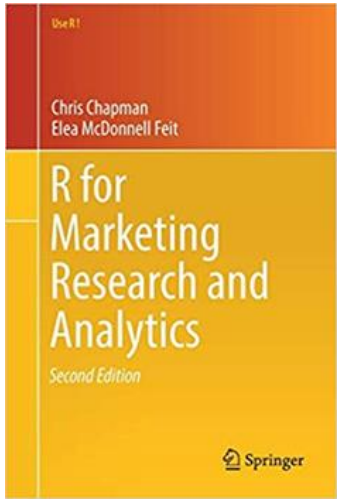
Winston, W. (2014). *Marketing Analytics: Data-Driven Techniques with Microsoft Excel.* Indianapolis, Indiana: John Wiley & Sons Inc.



Focuses on marketing analytics modeling techniques in Excel

This book is available at Amazon.com. The approximate cost should be approximately $27 - $32.

Chapman, C. and Feit, E. (2019). *R for Marketing Research and Analytics*. Switzerland AG: Springer Nature



Focuses on marketing analytics modeling techniques in the statistical programming language - ***R***

This book is available at Amazon.com. The approximate cost should be approximately $58.

What are Recommender Engines?

- Examples
- Do they matter?
- Netflix competition
- Steps
- Major types

Examples of Recommendation Engines[1]

Every Successful Product Or Business Has A Strong Recommendation Engine At Its Core.[1]

Amazon's — "Customers who bought this item also bought…".

Netflix's — "Other Movies You May Enjoy…"

Spotify's — "Recommended songs…"

Google's — "Visually Similar Images…"

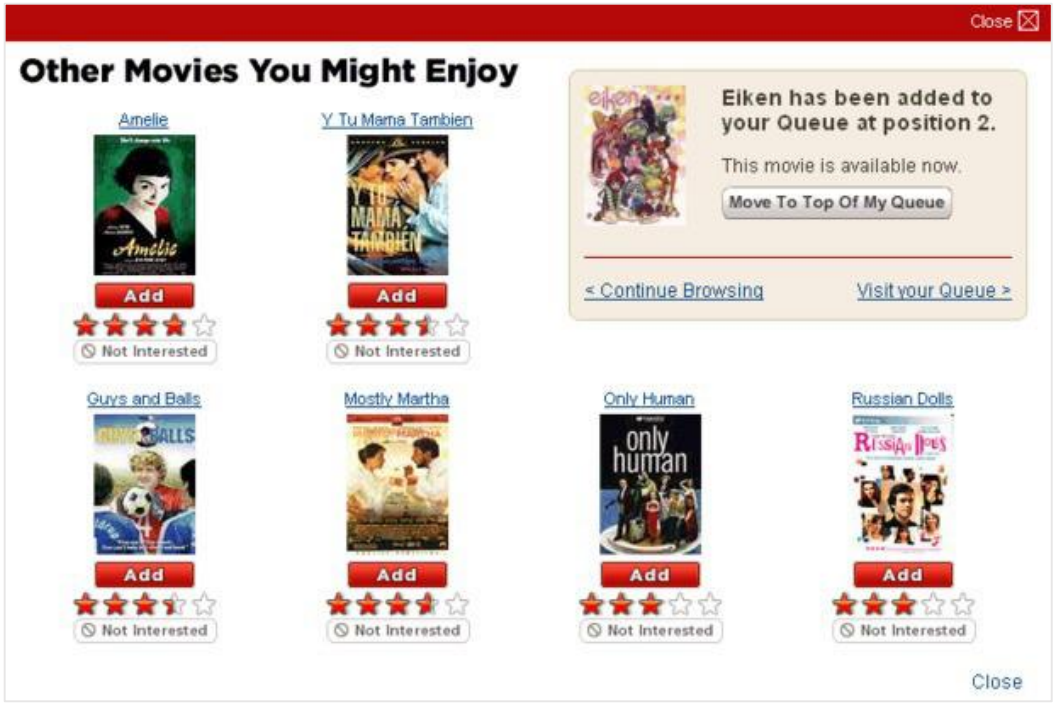
YouTube's — "Recommended Videos…"

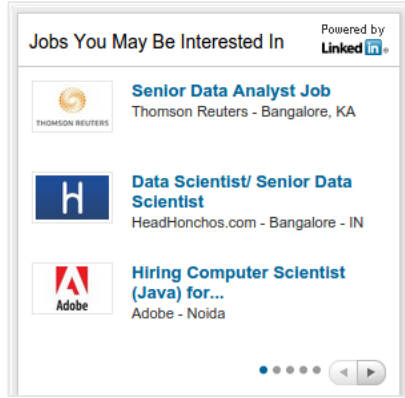
Facebook's — "People You May Know…"

LinkedIn's — "Jobs You May Be Interested In…"

Coursera's — "Recommended courses…" and

Waze's — "Best Route…"





1. Image captured from: How Can We Design An Intelligent Recommendation Engine - Hyperlink

Pictures of example recommendations copied from: An Introduction to Recommendation Engines – Hyperlink

Value of Recommendation Engines

- Chapter 10 in the required text discusses ***Predicting Individual Recommendations for Each Customer***
  - The text describes making individual recommendations as an approach to making a business more relevant to a customer
  - The text sells short the impact of recommendations on sales / utilization of a site
- Jannach, D., Jugovac, M. (2019). Measuring the Business Value of Recommender Systems. ACM Transactions on Management Information Systems, 10(4).[1]
  - Click Through Rates
  - Adoption and Conversion
  - Sales & Revenue
  - Effects on Sales Distributions
  - User Engagement & Behavior

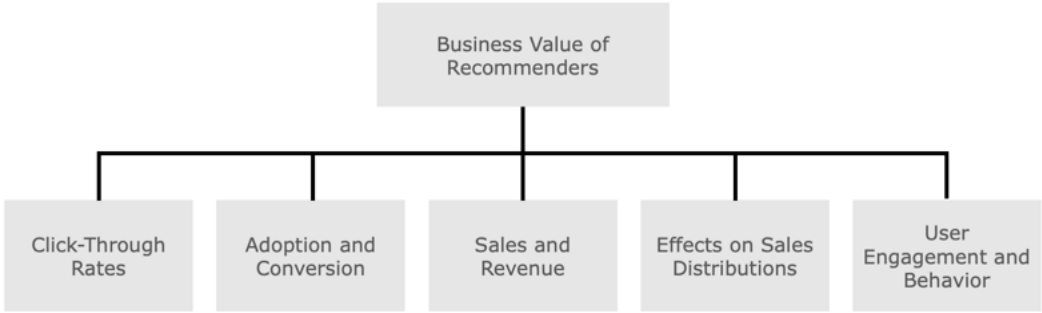


Fig. 1. Overview of Measurement Approaches

1. Image captured from: Jannach & Jogovac, Figure 1

Value of Recommendation Engines

- Observations of improvements from research:
  - Google News – Click Through Rate (CTR) improved by 38% when personalized recommendations were compared to only popular items
  - YouTube recommendations increased the CTR by 200% when personalized recommendations were compared with a “most-viewed” video recommendation
  - Adoption and conversion rates were tested with an A/B test at eBay
    - Bid-rates increased between 3.3 and 9% when recommendations were personalized
  - Sales and Revenue indicated a 35% improvement in sales compared with a “no-recommendation” condition on DVD sales from an online retailer
  - Profit improved by 28% at an online book store when using a personalized recommender engine for recommended books compared to a simple recommendation
    - Revenue dropped by 17% when recommendations were removed from the online book store’s website
  - Sales distributions (the range of different items purchased, in addition to the overall sales) expanded when using personalized recommendations
  - User behavior and engagement improved in media streaming which led to increased levels of user retention

1. Image captured from: Jannach & Jogovac, Figure 1

Value of Recommendation Engines

- There are challenges to measuring the value of recommendation engines
  - A/B tests
  - Accuracy (RMSE) as a measure of value?

Table 1. Measurements to Assess the Value of Recommenders.

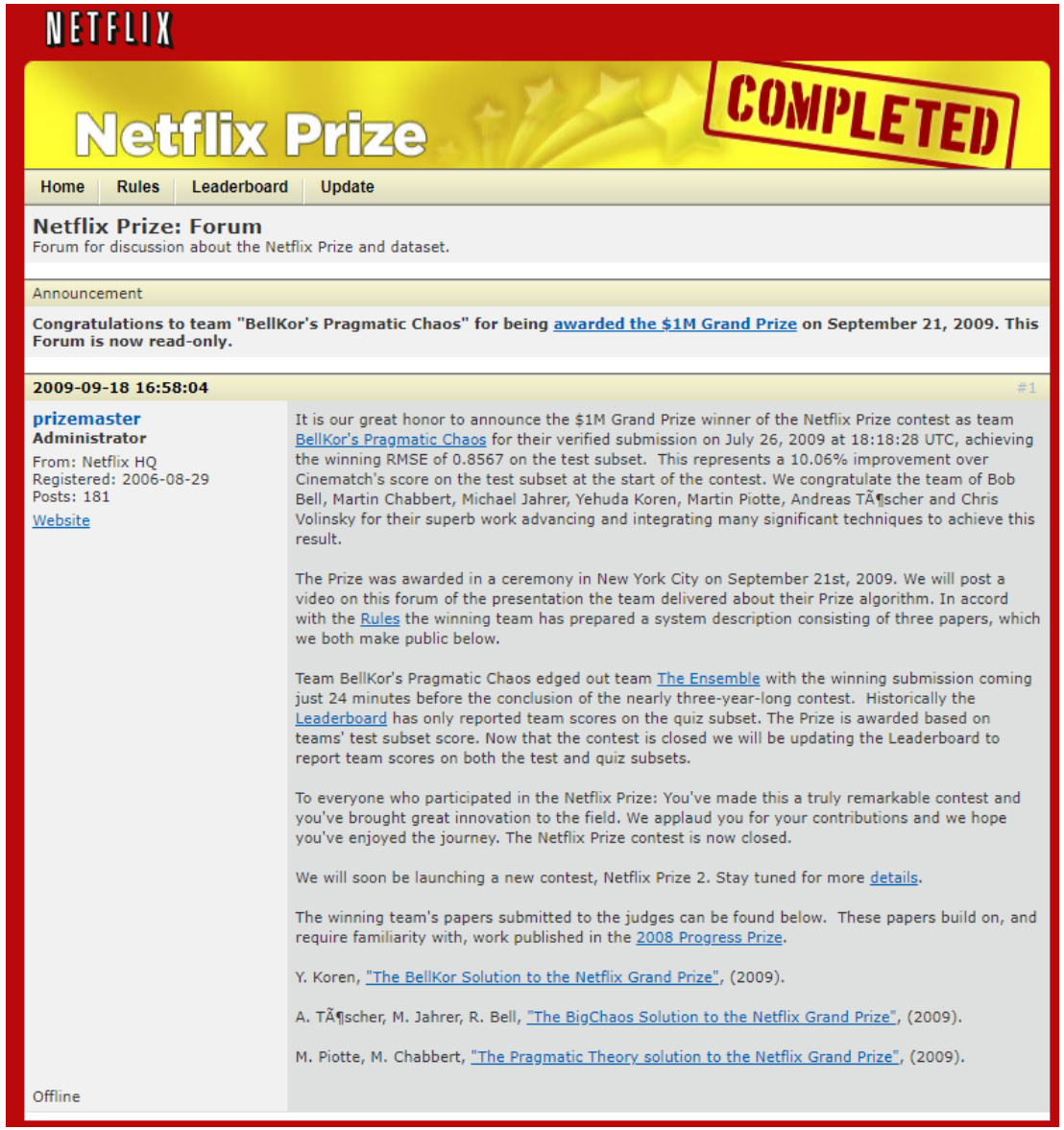
Measurement	Remarks
Click-Through Rates	Easy to measure and established, but often not the ultimate goal.
Adoption and Conversion	Easy to measure, but often requires a domain- and application specific definition. Requires interpretation and does not always translate directly into business value.
Sales and Revenue	Most informative measure, but cannot always be determined directly.
Effects on Sales Distribution	A very direct measurement; requires a thorough understanding of the effects of the shifts in sales distributions.
User Engagement and Behavior	Often, a correspondence between user engagement and customer retention is assumed; still, it remains an approximation.

1. Image captured from: Jannach & Jogovac, Table 1

Value of Recommendation Engines

- Jannach, D., Jugovac, M. (2019). Measuring the Business Value of Recommender Systems. ACM Transactions on Management Information Systems, 10(4).
  - Excellent article on how much business value the recommendations generate
- Excellent article on Item-based Collaborative Filtering Recommendation Algorithms
  - Sarwar, B., Karypis, G., Konstan, J., Reidl, J. (2001). Item-Based Collaborative Filtering Recommendation Algorithms. *Proceedings of the 10th International Conference on World Wide Web*, 285-295.
  - Download of PDF available at: https://dl.acm.org/doi/pdf/10.1145/371920.372071?download=true
- Toy example spreadsheet adapted from:
  - Winston, W. (2014). *Marketing Analytics: Data-Driven Techniques with Microsoft Excel* (pp. 393-402). Indianapolis, IN: Wiley and Sons.

The Netflix $1M Challenge



What Did Netflix Ask People to Do for $1M?

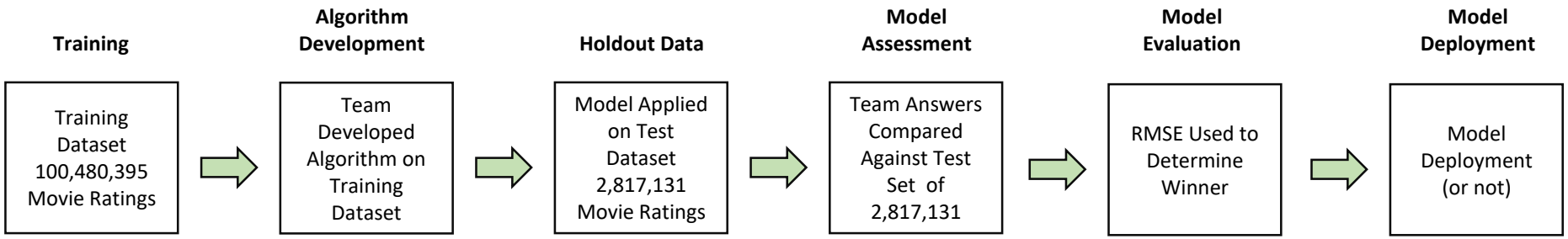
- Dataset – Over 100,000,000 ratings of 17,770 movies from 480,189 customers – Hyperlink on Kaggle
- Great Article that explains the competition – Hyperlink

In Netflix's own words:

- We're quite curious, really. To the tune of one million dollars.
- Netflix is all about connecting people to the movies they love. To help customers find those movies, we've developed our world-class movie recommendation system: CinematchSM. Its job is to predict whether someone will enjoy a movie based on how much they liked or disliked other movies. We use those predictions to make personal movie recommendations based on each customer's unique tastes. And while Cinematch is doing pretty well, it can always be made better.
- Now there are a lot of interesting alternative approaches to how Cinematch works that we haven't tried. Some are described in the literature, some aren't. We're curious whether any of these can beat Cinematch by making better predictions. Because, frankly, if there is a much better approach it could make a big difference to our customers and our business.
- So, we thought we'd make a contest out of finding the answer. It's "easy" really. We provide you with a lot of anonymous rating data, and a prediction accuracy bar that is 10% better than what Cinematch can do on the same training data set. (Accuracy is a measurement of how closely predicted ratings of movies match subsequent actual ratings.) If you develop a system that we judge most beats that bar on the qualifying test set we provide, you get serious money and the bragging rights. But (and you knew there would be a catch, right?) only if you share your method with us and describe to the world how you did it and why it works.
- Serious money demands a serious bar. We suspect the 10% improvement is pretty tough, but we also think there is a good chance it can be achieved. It may take months; it might take years. So to keep things interesting, in addition to the Grand Prize, we're also offering a $50,000 Progress Prize each year the contest runs. It goes to the team whose system we judge shows the most improvement over the previous year's best accuracy bar on the same qualifying test set. No improvement, no prize. And like the Grand Prize, to win you'll need to share your method with us and describe it for the world.
- There is no cost to enter, no purchase required, and you need not be a Netflix subscriber. So if you know (or want to learn) something about machine learning and recommendation systems, give it a shot. We could make it really worth your while.

Outcome of the Netflix Challenge

- Netflix Never Used Its $1M Algorithm Due to Engineering Costs – Hyperlink
  - Great example of the Deployment step of the Analytics Process



- De-anonomization by comparing Netflix ratings (which were private) with IMDB ratings which were public - Hyperlink

What is Collaborative Filtering?

- Collaborative filtering is an algorithmic approach to predicting what a customer would like given the customer's and previous choices
- Collaborative filtering is used by many recommendation engines
- There are two main types of collaborative filtering algorithmic approaches:
  - User-based
  - Item-based
- User-based is an approach that relates similarity between people to determine what to recommend
  - Example – Netflix determines that "users like me" have liked a move and recommend that movie to me
- Item-based is an approach that relates similarity between items to determine what to recommend
  - Example – Amazon determines that a user has read books on Data Science using R and recommends new books on Data Science using R when they are published

Examples of Recommendation Engines[1]

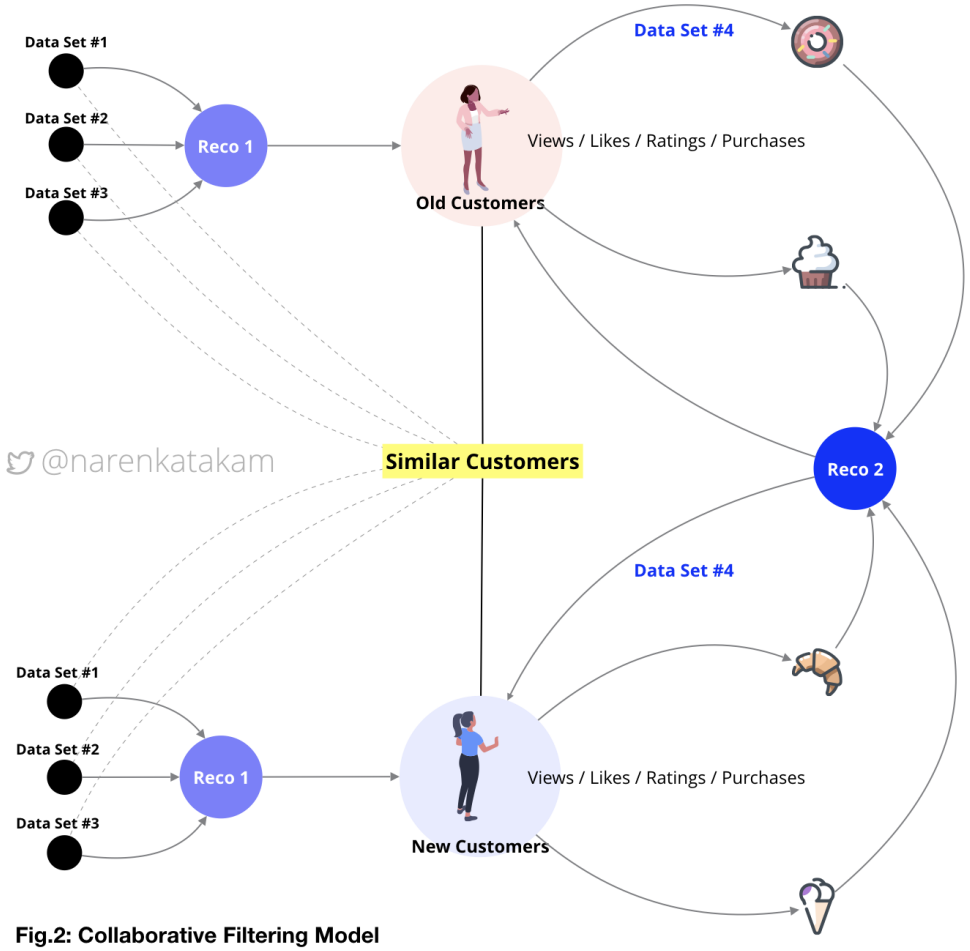


Fig.2: Collaborative Filtering Model

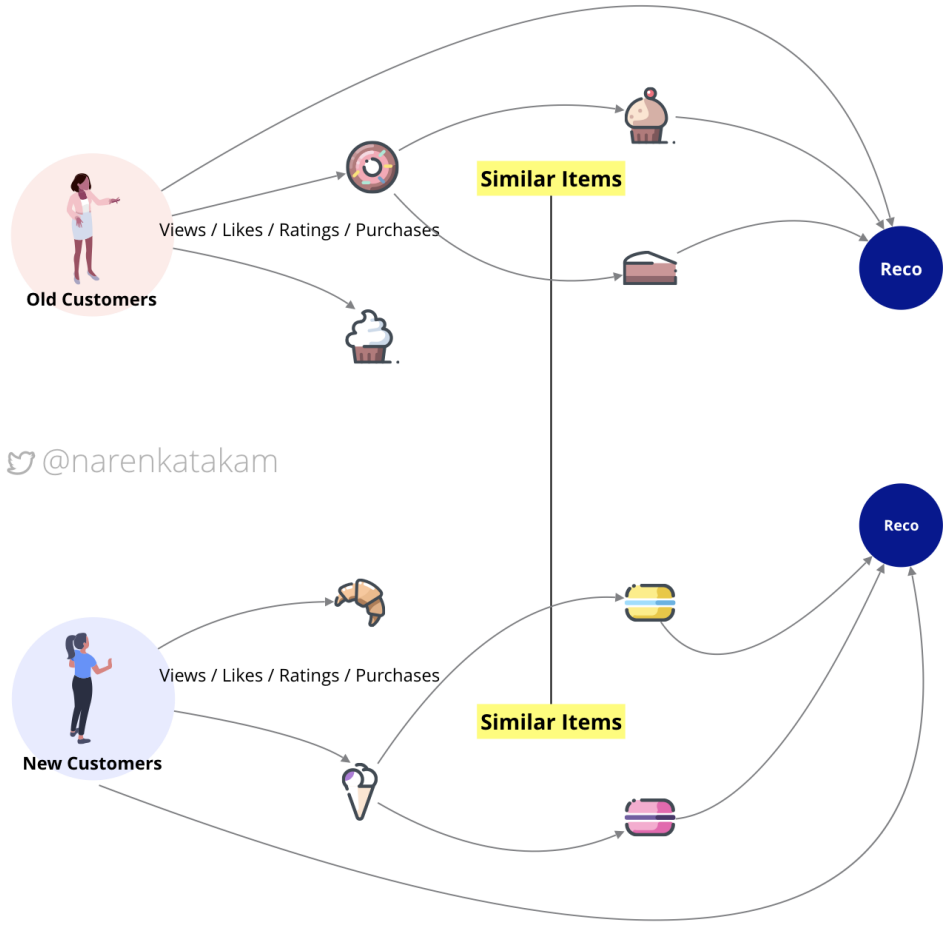


Fig.3: Content Filtering Model

User-based collaborative filtering

Item-based collaborative filtering

1. Images captured from: How Can We Design An Intelligent Recommendation Engine - Hyperlink

Let's Build A Recommendation Engine

- We will use the standard analytics methodology process
- The model is built in Excel
- The spreadsheet can be downloaded from: https://github.com/Hoover-code/DBA/blob/master/CollaborativeFilteringExample.xlsx

Analytics Process

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

User-Based Recommendation Engine Example – Business Understanding

- Business problem:
  "We are thinking about building a recommendation engine to replace Yelp for restaurants. Can you build me a couple of examples of recommendation engines for restaurants around UF?"
- Success definition:
  "I would define success for this project, if you could predict how much someone would like a restaurant based on how much others who are like the person of interest like the restaurant. That way, we could make restaurant recommendations to the person of interest."

CEO Business Request

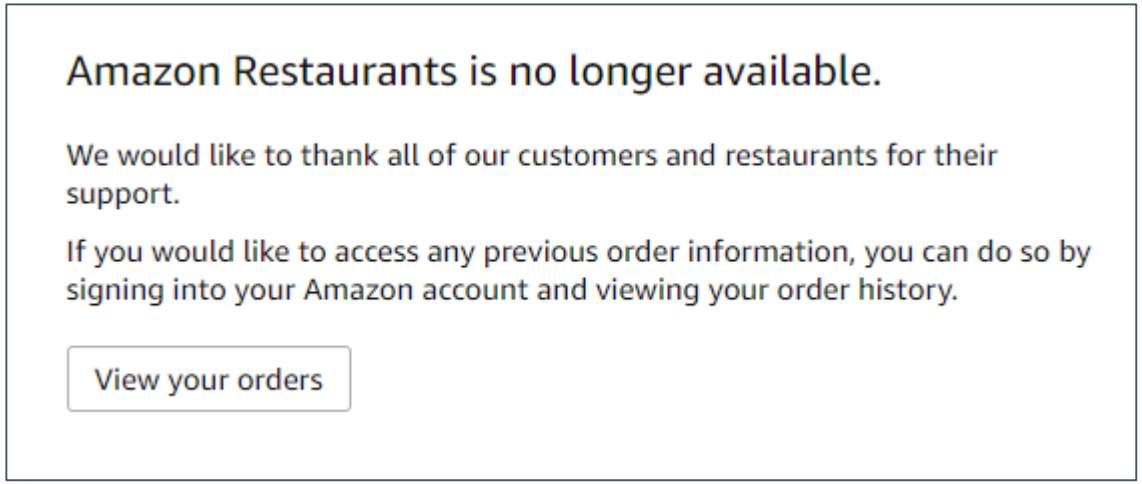


Image captured from: https://www.amazon.com/restaurants/legacy

Hyperlink to video about Amazon's recommendation engine

Hyperlink to reference article about Amazon Restaurants

Analytics Process

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

User-Based Recommendation Engine Example – Data Understanding

- Download the spreadsheet from GitHub: https://github.com/Hoover-code/MAR6930/blob/master/CollaborativeFilteringExample.xlsx
- Go to the RawData tab
- This is a "toy" dataset to illustrate building a recommendation engine
- The data was made up for illustration purposes
- Seven **made-up** members of the UF Warrington Marketing Department rated six different restaurants within 0.25 miles of the business school
- The ratings were on a scale of 1-5 with one being least preferred and 5 being most preferred
- Names and ratings are for illustrative purposes only – names were changed to protect the innocent!

	Restaurant Preferences					
Group Members	Chipotle	Sushi2Go	Chick-Fil-A	Just Salad	Firehouse Subs	Rising Roll
Alan Cooks	5	2	5	3	4	2
Rich Luster	2	3	4	1	5	1
Aner Celebrate	3	4	3	5	3	2
Yang Tree	4	1	3	5	2	3
Steve Tough	3	3	5	1	4	1
Jim Vacuum	4	3	4	1	3	1
Sian French	2	4	3	5	2	2

Analytics Process

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

User-Based Recommendation Engine Example – Data Preparation

- Because this is a “Toy” dataset, data preparation is not required

	Restaurant Preferences					
Group Members	Chipotle	Sushi2Go	Chick-Fil-A	Just Salad	Firehouse Subs	Rising Roll
Alan Cooks	5	2	5	3	4	2
Rich Luster	2	3	4	1	5	1
Aner Celebrate	3	4	3	5	3	2
Yang Tree	4	1	3	5	2	3
Steve Tough	3	3	5	1	4	1
Jim Vacuum	4	3	4	1	3	1
Sian French	2	4	3	5	2	2

Analytics Process

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

User-Based Recommendation Engine Example – Modeling

- The modeling approach is going to predict Jim Vacuum's rating for a restaurant he hasn't been to – Rising Roll
- To begin with, we will calculate the mean for Jim Vacuum's average rating for all of the restaurants he has rated
- Then we will identify the people whose ratings on restaurants are the most like the ratings from Jim Vacuum for the same restaurants
- We will adjust the ratings of each person who has rated Rising Roll to adjust Jim Vacuum's average rating. The more similar the person's other ratings are to Jim Vacuum's, the more weight that we will give their ratings.
- There are several ways to evaluate similarity of ratings:
  - Correlation between the ratings of other people with Jim Vacuum is one way
  - If the correlation between a person and Jim Vacuum is close to +1, then if the other person likes a restaurant, Jim Vacuum is more likely to like the restaurant too
  - If the correlation between a person and Jim Vacuum is close to -1, the if the other person likes a restaurant, Jim Vacuum is less likely to like that restaurant
  - The Excel CORREL function can determine the correlation between two data sets

Analytics Process

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

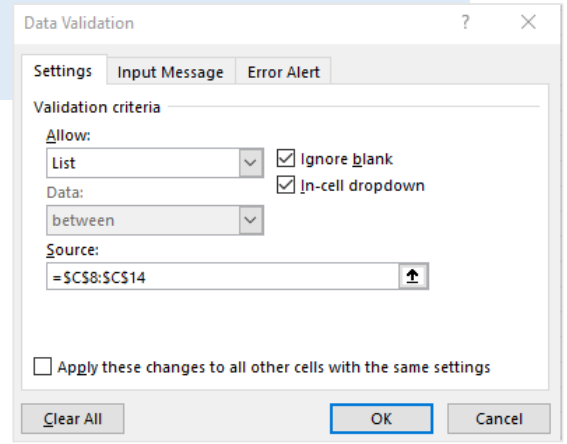
User-Based Recommendation Engine Example – Modeling

- Download the spreadsheet from this location: https://github.com/Hoover-code/DBA/blob/master/CollaborativeFilteringExample.xlsx
- Go to the **Build** tab
- Let's calculate the correlation between all of the restaurant raters
- First, calculate the mean of each rater in Column J
  - In Cell J7, enter the title Mean
  - In Cell J8, enter the formula: =AVERAGE(D8:I8)
  - Then copy the formula from J8:J14
- In Row 15, we are going to enter an index number
  - The index number just illustrates which column in the table is the one that we will "point to"
  - In Cells D15:i15, enter the numbers 1-6 respectfully
- In Cells C16 and C17, we will enter 2 of the names of the raters to compare with a correlation
  - Let's learn to use a data validation drop down box to select the names exactly as they are in the table above
  - First, enter the title Person 1 in cell B16
  - Next, enter the title Person 2 in cell B17

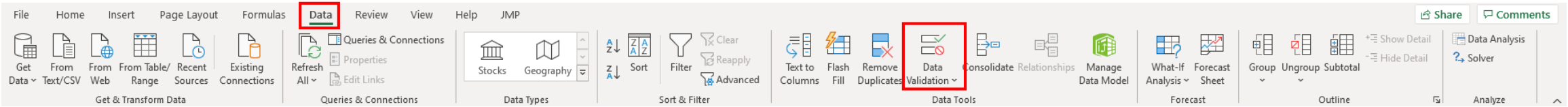
Analytics Process

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

User-Based Recommendation Engine Example – Modeling



- Here is how to use the data validation to create a drop-down box
  - Place your cursor in Cell C16
  - Go to the Data selection on the Menu bar
  - Then select Data Validation on the Ribbon bar
  - Select List from the validation menu



  - Then select the range under source: =$C$8:$C$14
  - Click Ok
  - Now when you put your cursor in Cell C16, you can select one of the names in Cells C8:C14 exactly
  - Select Rich Luster with the drop-down box
  - Complete the same steps for cell C17
  - Select Jim Vacuum with the drop-down box

Analytics Process

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

User-Based Recommendation Engine Example – Modeling

- Now let's use rows 16 and 17 to calculate the correlation between the ratings of the two folks that we randomly selected
  - In Cell D16, enter the formula: =INDEX($D$8:$I$14,MATCH($C16,$C$8:$C$14,0),D$15)
  - This formula captures the data that is in the ratings table above and matches Person 1 with the person in table with the same Exact name
  - Copy that formula from cell D16:i16
  - Copy that formula to cell D17:i17
  - Validate that it is looking up in the table correctly

Hyperlink to a reference article about the Excel Index Function

Hyperlink to a reference article about the Excel Match Function

		Restaurant Preferences						
	Group Members	Chipotle	Sushi2Go	Chick-Fil-A	Just Salad	Firehouse Subs	Rising Roll	Mean
	Alan Cooks	5	2	5	3	4	2	3.5
	Rich Luster	2	3	4	1	5	1	2.666667
	Aner Celebrate	3	4	3	5	3	2	3.333333
	Yang Tree	4	1	3	5	2	3	3
	Steve Tough	3	3	5	1	4	1	2.833333
	Jim Vacuum	4	3	4	1	3		3
	Sian French	2	4	3	5	2	2	3
		1	2	3	4	5	6	
Person 1	Rich Luster	=INDEX($D$8:$I$14,MATCH($C16,$C$8:$C$14,0),D$15)				5	1	
Person 2	Jim Vacuum	4	3	4	1	3	0	

User-Based Recommendation Engine Example – Modeling

- We can't calculate correlations on missing values (cell i13)
  - We need to ensure that only cells with completed entries are correlated
- So we will add a couple of rows to error correct the missing value with a "_"
  - In Cell C18, enter =C16
  - In Cell C19, enter =C17
  - In Cell D18, enter the formula: =IF(COUNTIF(D$16:D$17,">0")=2,D16,"_ ")
  - Copy that formula to D18:i18
  - Copy that formula to D19:i19
- Note that for Rising Roll, there **will not** be a comparison

		Restaurant Preferences						
	Group Members	Chipotle	Sushi2Go	Chick-Fil-A	Just Salad	Firehouse Subs	Rising Roll	Mean
	Alan Cooks	5	2	5	3	4	2	3.5
	Rich Luster	2	3	4	1	5	1	2.666667
	Aner Celebrate	3	4	3	5	3	2	3.333333
	Yang Tree	4	1	3	5	2	3	3
	Steve Tough	3	3	5	1	4	1	2.833333
	Jim Vacuum	4	3	4	1	3		3
	Sian French	2	4	3	5	2	2	3
		1	2	3	4	5	6	
Person 1	Rich Luster	2	3	4	1	5	1	
Person 2	Jim Vacuum	4	3	4	1	3	0	
	Rich Luster	2	3	4	1	5	=IF(COUNTIF(I$16:I$17,">0")=2,I16,"_ ")	
	Jim Vacuum	4	3	4	1	3	_	

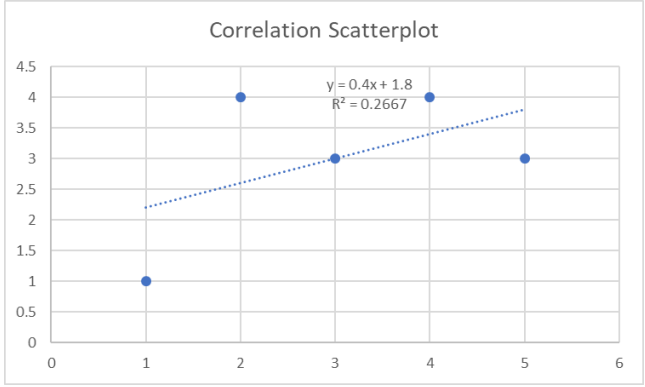
Rising Roll
2
1
2
3
1
2
6
1
0
_
_

Analytics Process

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

User-Based Recommendation Engine Example – Modeling

- Now let's calculate the correlation of the two raters
  - First, calculate the mean of the two raters in column J
    - In Cell J18, enter the formula: =AVERAGE(D18:I18)
    - Copy that formula to cell J19
    - The mean of both raters for the restaurants that both rated is 3.000
    - That seems very similar
  - Now let's calculate the correlation of the two raters
    - In Cell M19, enter the formula: =CORREL(D18:I18,D19:I19)
    - In Cell N19, enter the title Correlation
    - The Correlation is 0.516398
    - What can you say about how "alike" (correlated) these raters are?



Person 1	Rich Luster	2	3	4	1	5	1						
Person 2	Jim Vacuum	4	3	4	1	3	0						
	Rich Luster	2	3	4	1	5	_	3					
	Jim Vacuum	4	3	4	1	3	_	3				0.516398	Correlation

User-Based Recommendation Engine Example – Modeling

- Now let's Calculate the correlation between all pairs of raters
  - In Cell C22, enter the formula =CORREL(D18:I18,D19:I19)
  - This copies the correlation between the two people on Rows 18 & 19
- Next copy the names from cells C8:C14 to cells C23:C29
- Then, using the paste special, transpose capability, paste the names in cells D22:J22
- The table should now look like this

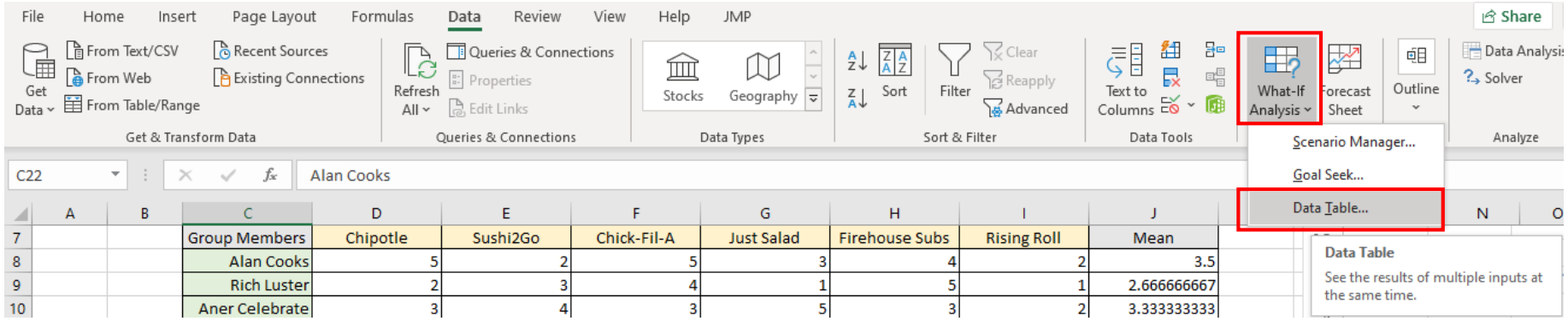
0.516397779	Alan Cooks	Rich Luster	Aner Celebrate	Yang Tree	Steve Tough	Jim Vacuum	Sian French
Alan Cooks							
Rich Luster							
Aner Celebrate							
Yang Tree							
Steve Tough							
Jim Vacuum							
Sian French							

Analytics Process

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

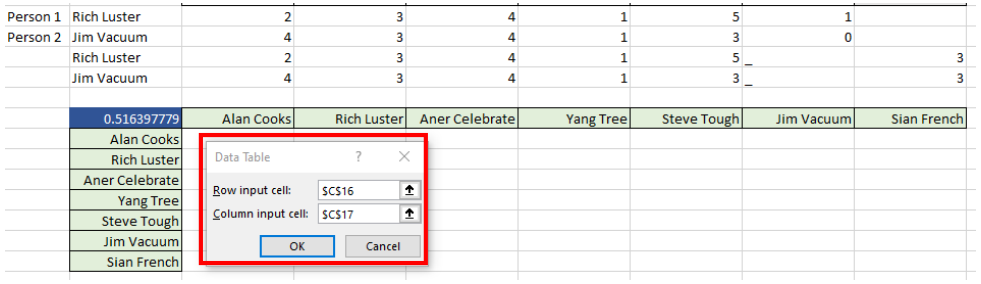
User-Based Recommendation Engine Example – Modeling

- We are now going to use a new capability from Excel – The "What-If Analysis" capability under the Data Menu



Group Members	Chipotle	Sushi2Go	Chick-Fil-A	Just Salad	Firehouse Subs	Rising Roll	Mean
Alan Cooks	5	2	5	3	4	2	3.5
Rich Luster	2	3	4	1	5	1	2.666666667
Aner Celebrate	3	4	3	5	3	2	3.333333333

- Select cells C22:J28, then select from the menu bar Data, What-if Analysis, Data Table
- Click on the data table, and a pop-up box will appear that asks for the row input cell and the column input cell
- This will "iterate" among the various values of Person 1 and Person 2 when you choose cells $C$16 for the Row input cell and $C$17 for the Column input cell



Hyperlink to a reference article about the Excel Data Table capability under What-If Analysis

Hyperlink to video about Data Tables under What-If Analysis

Analytics Process

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

User-Based Recommendation Engine Example – Modeling

- Now press OK
- Your output should look like a correlation matrix table

0.516397779	Alan Cooks	Rich Luster	Aner Celebrate	Yang Tree	Steve Tough	Jim Vacuum	Sian French
Alan Cooks	1	0.444261658	-0.140487872	0.307793506	0.679250119	0.626224291	-0.344123601
Rich Luster	0.444261658	1	-0.158113883	-0.606217783	0.891882585	0.516397779	-0.290473751
Aner Celebrate	-0.140487872	-0.158113883	1	0.273861279	-0.201455741	-0.912870929	0.918558654
Yang Tree	0.307793506	-0.606217783	0.273861279	1	-0.441367415	-0.387298335	0.223606798
Steve Tough	0.679250119	0.891882585	-0.201455741	-0.441367415	1	0.825722824	-0.296078263
Jim Vacuum	0.626224291	0.516397779	-0.912870929	-0.387298335	0.825722824	1	-0.782780364
Sian French	-0.344123601	-0.290473751	0.918558654	0.223606798	-0.296078263	-0.782780364	1

- Now that we can compare how closely raters compare to each other, we are ready to set up the prediction engine

Analytics Process

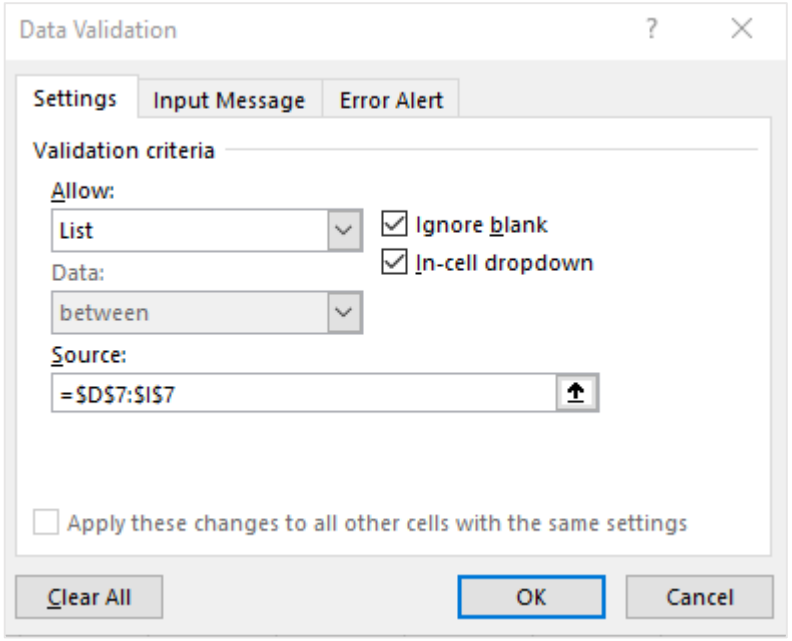
- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

User-Based Recommendation Engine Example – Modeling

- To set up the prediction engine, we need to identify the person that we are trying to predict a rating for the restaurant of interest
  - In Cell C31, enter the title Predict Rating for
  - In Cell D31, enter the title Person
  - We will use the Data Validation Approach to pick both of these entries

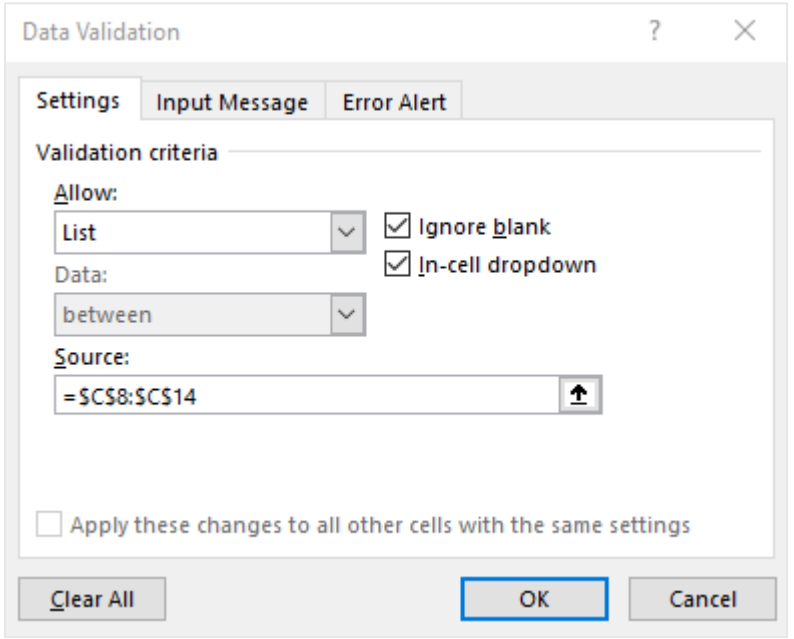
Predict Rating for	Person
Rising Roll	Jim Vacuum

In cell C32, set the data validation box like this



Then use the drop-down to select Rising Roll

In cell D32, set the data validation box like this



Then use the drop-down to select Jim Vacuum

User-Based Recommendation Engine Example – Modeling

- Use Vlookup to capture the mean for values
  - In cell E31, enter the title Mean
  - In Cell E32, enter the formula: =VLOOKUP(D32,$C$8:$J$14,8,FALSE)
  - In Cell F32, enter the formula: =C8 and copy from F32:F38
  - In Cell G31, enter the title Mean
  - In Cell G32, enter the formula: =VLOOKUP(F32,$C$8:$J$14,8,FALSE)
  - Copy that formula from G32:G38

Predict Rating for	Person	Mean		Mean
Rising Roll	Jim Vacuum	3	Alan Cooks	3.5
			Rich Luster	2.666666667
			Aner Celebrate	3.333333333
			Yang Tree	3
			Steve Tough	2.833333333
			Jim Vacuum	3
			Sian French	3

Analytics Process

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

User-Based Recommendation Engine Example – Modeling

- Now let's determine the similarity of the selected person to the other raters with by capturing the correlations from the table we created above
  - In Cell H31, enter the title Similarity
  - In Cell H32, enter the formula: =INDEX($D$23:$J$29,MATCH($D$32,$C$23:$C$29,0),MATCH(F32,$D$22:$J$22,0))
  - Copy that formula from Cell H32:H38
- Now, we will capture the restaurant rating for the restaurant we are predicting across all of the other raters
  - In Cell i31, enter the title Rest. Rating
  - In Cell i32, enter the formula: =INDEX($D$8:$I$14,MATCH(F32,$F$32:$F$38,0),MATCH($C$32,$D$7:$I$7,0))
  - Copy the formula from i32:i38
- Now let's adjust the rating between each rater and the person of interest
  - In Cell J31, enter the title Adjustment
  - In Cell J32, enter the formula: =IF(AND(F32<>$D$32,I32>0),(I32-G32),0)
  - Copy that formula from J32:J38

Analytics Process

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

User-Based Recommendation Engine Example – Modeling

- Your spreadsheet should look like this

Predict Rating for	Person	Mean		Mean	Similarity	Rest. Rating	Adjustment
Rising Roll	Jim Vacuum	3	Alan Cooks	3.5	0.626224291	2	=IF(AND(F32<>$D$32,I32>0),(I32-G32),0)
			Rich Luster	2.666666667	0.516397779	1	-1.666666667
			Aner Celebrate	3.333333333	-0.912870929	2	-1.333333333
			Yang Tree	3	-0.387298335	3	0
			Steve Tough	2.833333333	0.825722824	1	-1.833333333
			Jim Vacuum	3	1	0	0
			Sian French	3	-0.782780364	2	-1

  - This adjustment modifies the mean rating across all restaurants by rater to the specific rating for the restaurant of interest (i.e. Rising Roll)
- Now we will make an abosolute value adjustment column that applies an absolute value to the similarity value
  - In Cell K31, enter the title Abs. Correlation
  - In Cell K32, enter the formula: =IF(AND(F32<>$D$32,I32>0),ABS(H32),0)
  - Copy that formula from cell K32:K38
- We are almost ready to make our prediction!

User-Based Recommendation Engine Example – Modeling

- Here are the final calculations for the prediction:
  - In Cell F40, enter the title Total Adjustment
  - In Cell F41, enter the title Final Rating – this is the prediction!
- For the Total Adjustment, we will apply the Excel Sumproduct function
  - In Cell G40, enter the formula: =SUMPRODUCT(J32:J38,H32:H38)/SUM(K32:K38)
- Now we will adjust the final predicted rating for Jim Vacuum's at the Rising Roll
  - In Cell G41, enter the formula: =E32+G40

Predict Rating for	Person	Mean		Mean	Similarity	Rest. Rating	Adjustment	Abs. Correlation
Rising Roll	Jim Vacuum	3	Alan Cooks	3.5	0.626224291	2	-1.5	0.626224291
			Rich Luster	2.666666667	0.516397779	1	-1.666666667	0.516397779
			Aner Celebrate	3.333333333	-0.912870929	2	-1.333333333	0.912870929
			Yang Tree	3	-0.387298335	3	0	0.387298335
			Steve Tough	2.833333333	0.825722824	1	-1.833333333	0.825722824
			Jim Vacuum	3	1	0	0	0
			Sian French	3	-0.782780364	2	-1	0.782780364
			Total Adjustment	=SUMPRODUCT(J32:J38,H32:H38)/SUM(K32:K38)				
			Final Rating	2.675688112				

Analytics Process

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

User-Based Recommendation Engine Example – Modeling

- Here is the final outcome:
  - The final adjusted rating is: 2.6757

Predict Rating for	Person	Mean		Mean	Similarity	Rest. Rating	Adjustment	Abs. Correlation
Rising Roll	Jim Vacuum	3	Alan Cooks	3.5	0.626224291	2	-1.5	0.626224291
			Rich Luster	2.666666667	0.516397779	1	-1.666666667	0.516397779
			Aner Celebrate	3.333333333	-0.912870929	2	-1.333333333	0.912870929
			Yang Tree	3	-0.387298335	3	0	0.387298335
			Steve Tough	2.833333333	0.825722824	1	-1.833333333	0.825722824
			Jim Vacuum	3	1	0	0	0
			Sian French	3	-0.782780364	2	-1	0.782780364
			Total Adjustment	-0.324311888				
			Final Rating	2.675688112				

- What does that mean?
- Jim Vacuum's predicted rating for the Rising Roll restaurant would be 2.67 using the user-based recommendation engine

Analytics Process

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

User-Based Recommendation Engine Example – Modeling

- Here what all of the spreadsheet torture is doing

Estimate of Jim Vacuum's Rating for Rising Roll = (Jim's Mean Rating) +

$$\frac{\sum_{Oher\ Raters}(Similarity\ of\ Raters\ to\ Jim) * (Other\ Raters'rating\ for\ Rising\ Roll\ - Other\ Raters'average\ rating)}{\sum_{All\ Raters}|Other\ Raters'similarity\ to\ Jim|}$$

Analytics Process

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

User-Based Recommendation Engine Example – Modeling

- Let's practice predicting!
- Using the spreadsheet FoodUserBased.xlsx
  - Predict the score for Yang Tree at Sushi2Go
    - Final value should be 2.8521
  - Predict the score for Rich Luster at Rising Roll
    - Final value should be 1.9137
  - Predict the score for Aner Celebrate at Chick-Fil-A
    - Final value should be 2.6536
- Class Exercise – Let's join up in teams and practice changing the values
  - Each team member enter your rating in the matrix for each restaurant
  - Then choose one restaurant / person combination and see how close the final predicted rating matches the actual rating

Analytics Process

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

User-Based Recommendation Engine Example – Modeling Assessment

- How do these values compare with their actual rating?
  - Aner Celebrate at Chick-Fil-A actual score is 3.0
  - Aner Celebrates predicted score is 2.6536
  - Is that good?

- Root Mean Square Error (RMSE) is a way to evaluate predictiveness
- See the Netflix information at the beginning of the presentation

$$RMSE = \sqrt{\sum all\ ratings\ in\ the\ test\ set \frac{(actual\ rating\ - predicted\ rating)^2}{N}}$$

Analytics Process

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

Item-based Recommendation Engine Example - Modeling

- Another way we could have done the recommendation engine is to use an item-based approach
  - In this case, item would equal the restaurant
- This exercise has been completely captured with an item-based approach in the spreadsheet on Canvas: FoodItemBased.xlsx
- Open this spreadsheet and let's try changing the person / restaurant mix
  - Note, the correlation matrix is comparing restaurant to restaurant
  - That comparison indicates how different each restaurant is from the other restaurants
- Change the person in cell C34
- Change the restaurant in cell C35
- Read out the Final Rating Estimate in Cell D35
  - Try Rich Luster and Rising Roll – 1.2407
  - Yang Tree at Sushi2Go – 1.7764
  - Aner Celebrate at Chick-Fil-A – 3.0010

Analytics Process

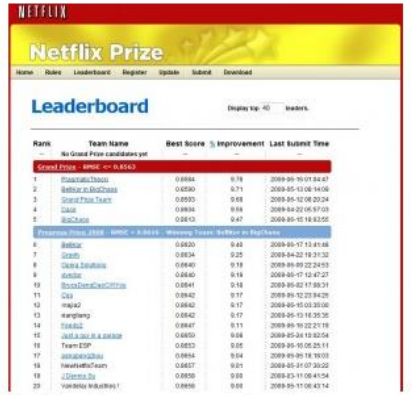
- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment

Item-based Recommendation Engine Example – Evaluation & Deployment

- Our models seem to work pretty well
- We have met the CEO's challenge
- However, it is a "Toy" model
  - Will this model scale to production?
- Deployment will require a significant IT investment
  - Note what happened on the Netflix challenge
  - Total new model not deployed due to Engineering Costs of implementing the complex algorithm

Hyperlink to a reference article about the reasons that Netflix did not deploy the algorithm

Netflix Never Used Its $1 Million Algorithm Due To Engineering Costs



Netflix awarded a $1 million prize to a developer team in 2009 for an algorithm that increased the accuracy of the company's recommendation engine by 10 percent. But it doesn't use the million-dollar code, and has no plans to implement it in the future, Netflix announced on its blog Friday. The post goes on to explain why: a combination of too much engineering effort for the results, and a