XII международная конференция

CEE-SECR / РАЗРАБОТКА ПО

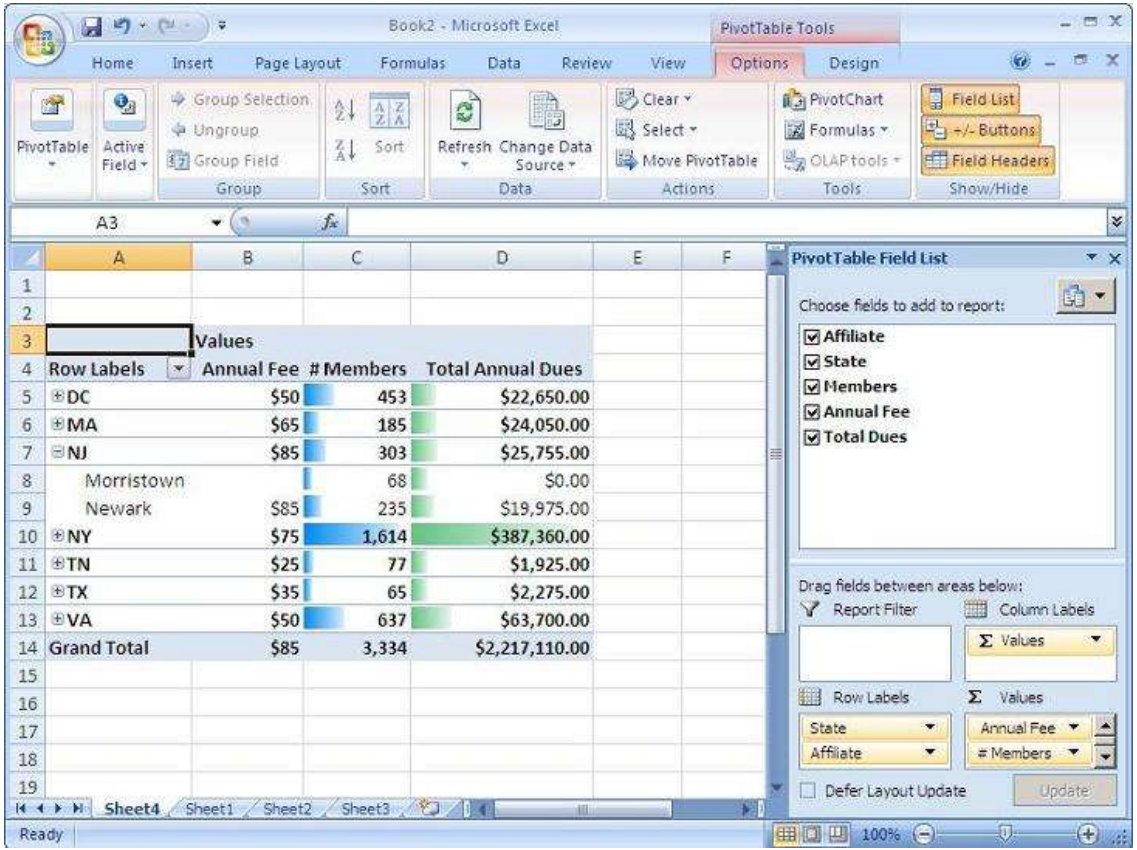
28 - 29 октября, Москва

Машинное обучение на каждый день

Евгений Виноградов

Яндекс.Деньги

BI


Book2 - Microsoft Excel
PivotTable Tools
Home
Insert
Page Layout
Formulas
Data
Review
View
Options
Design
PivotTable
Active Field
Group Selection
Ungroup
Group Field
Group
Sort
Sort
Refresh
Change Data Source
Data
Clear
Select
Move PivotTable
Actions
PivotChart
Formulas
OLAP tools
Tools
Field List
+/- Buttons
Field Headers
Show/Hide
A3
Values
Row Labels
Annual Fee
Members
Total Annual Dues
DC $50 453 $22,650.00
MA $65 185 $24,050.00
NJ $85 303 $25,755.00
Morristown 68 $0.00
Newark $85 235 $19,975.00
NY $75 1,614 $387,360.00
TN $25 77 $1,925.00
TX $35 65 $2,275.00
VA $50 637 $63,700.00
Grand Total $85 3,334 $2,217,110.00
Sheet4
Sheet1
Sheet2
Sheet3
Ready
PivotTable Field List
Choose fields to add to report:
Affiliate
State
Members
Annual Fee
Total Dues
Drag fields between areas below:
Report Filter
Column Labels
Values
Row Labels
Values
State
Affiliate
Annual Fee
Members
Defer Layout Update
Update
100%

А дальше?

Что такое машинное обучение

- Машинное обучение (англ. Machine Learning) — обширный подраздел искусственного интеллекта, математическая дисциплина, использующая разделы математической статистики, численных методов оптимизации, теории вероятностей, дискретного анализа, и извлекающая знания из данных.

Разделение труда

Разделение труда

- Data Scientist vs. Предметная область

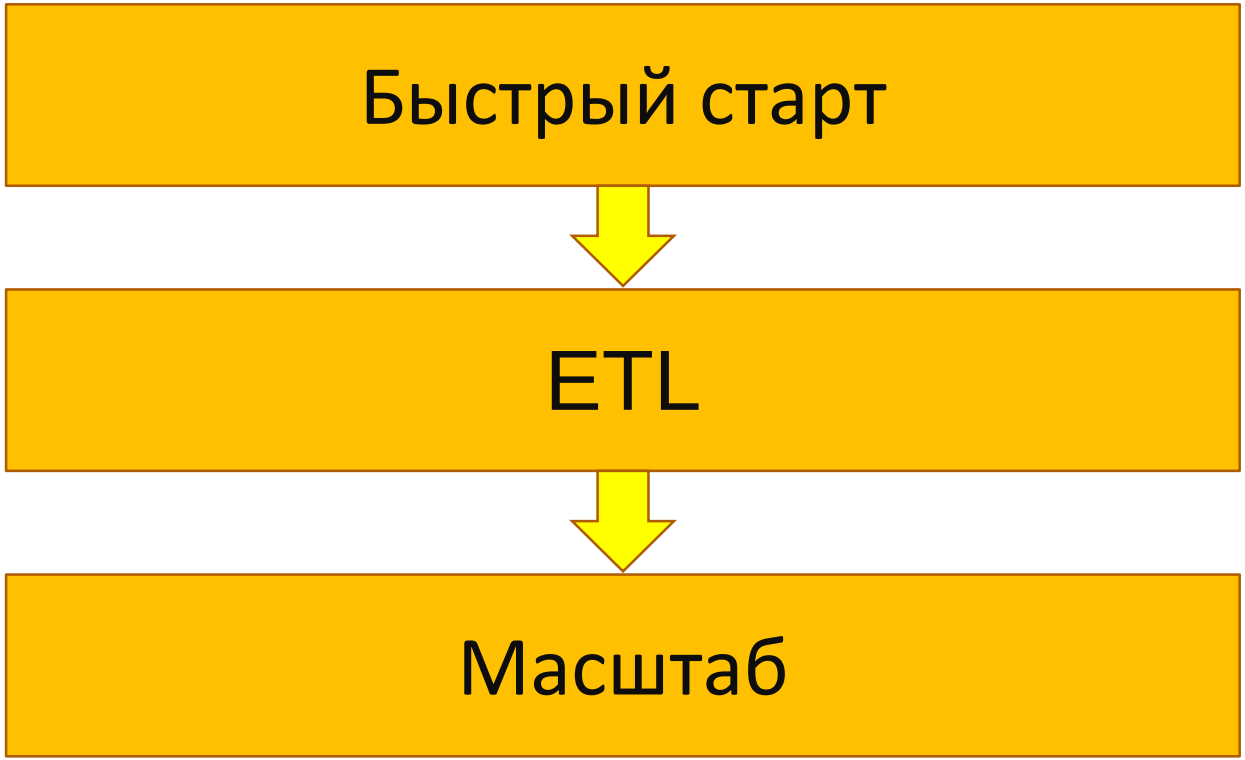
Внедрение


КАК НАРИСОВАТЬ СОВУ
1.
2.
РИСУЕМ КРУЖОЧКИ
РИСУЕМ ОСТАТОК СОВЫ
joyreactor.cc

Кружочки

- Экосистема
- Продвижение
- Поддержка

Экосистема



Продвижение

- Мифы
- Выбор задачи
- Сложность

Мифы

- Серебряная пуля
- Недооценка

Когда это не работает

- Зависимость в данных отсутствует

Загадка.

Самолет летит из Москвы в Нью-Йорк. Расстояние составляет три тысячи километров. Высота полета составляет одиннадцать тысяч четыреста сорок метров. На борту 296 пассажиров.
Сколько лет пилоту?

Когда это не работает

- Зависимость в данных отсутствует
- Цена ошибки выше оценки точности

Когда это не работает

- Зависимость в данных отсутствует
- Цена ошибки выше оценки точности
- Проблема в мнениях

Сложность

- Выбор инструмента
- Коллаборация
- Евангелисты
- Это RND

Пример задачи

- Наркотики
- Оружие
- Порнография
- Пропаганда насилия

Запрещённые категории товаров и услуг

- Пищевые добавки
- Товары для взрослых
- Торрент-трекеры
- Копии известных брендов

Категории товаров с высоким риском

Business Risk Assessment and Mitigation (BRAM) для MasterCard
Global Brand Protection Program (GBPP) для Visa

Пред- и пост-обработка

- Автоматическое скачивание сайтов
- Ручное создание обучающего набора данных
- Неопределённость некоторых категорий
- Уточнение результатов классификации

Обучающий процесс

TF-IDF-матрица

label	metadata_file	ноутбук	ноч	ночн	ночник
normal	1000pechi.ru.txt	0	0	0	0
normal	10kr.ru.txt	0	0	0	0
normal	1popov.ru.txt	0.003	0.003	0	0
adult	24poppers.ru.txt	0	0.004	0.001	0
normal	28panfilovcev.com.txt	0	0	0	0
normal	3sunduka.ru.txt	0.040	0.003	0.003	0
supplements	5lb.ru.txt	0	0	0	0
normal	5plusov.ru.txt	0.003	0	0	0
replica	64-64.ru.txt	0	0	0	0
drugs	AlcoMag.ru.txt	0	0	0.020	0
drugs	VinAlco.ru.txt	0	0	0	0

Результат

	true adult	true drugs	true replica	true weapons	true normal	true betting exchange	true hourhotel	true magic	true spy	true supplements	true torrent	**class precision**
pred. **adult**	14	1	0	0	4	0	0	0	0	0	0	**73.68%**
pred. **drugs**	1	10	0	0	0	0	0	0	0	0	0	**90.91%**
pred. **replica**	1	0	12	0	1	0	0	0	0	0	0	**85.71%**
pred. **weapons**	0	0	0	10	0	0	0	0	0	0	0	**100.00%**
pred. **normal**	1	2	1	0	88	4	0	1	0	0	1	**89.80%**
pred. **betting exchange**	0	0	0	0	0	9	0	0	0	0	0	**100.00%**
pred. **hourhotel**	0	0	0	0	2	0	8	0	0	0	0	**80.00%**
pred. **magic**	0	0	0	0	1	0	0	8	0	0	0	**88.89%**
pred. **spy**	0	0	0	0	1	0	0	0	9	0	0	**90.00%**
pred. **supplements**	0	0	0	0	0	0	0	0	0	12	0	**100.00%**
pred. **torrent**	0	0	0	0	2	0	0	0	0	0	4	**66.67%**
class recall	**82.35%**	**76.92 %**	**92.31%**	**100.00%**	**88.89%**	**69.23%**	**100.00%**	**88.89%**	**100.00%**	**100.00%**	**80.00 %**	

Key takeaways

- Экосистема
- Обучение
- Истории успеха

Key takeaways

- Математика никуда не пропадает

Вопросы?

Евгений Виноградов, Яндекс.Деньги

- jonny@yamoney.ru
- @evinogradov