Intern @ Slice Intelligence

Weitian Wu

September 8, 2014

slice

Outline

- Details about the job
- Skills required and learned
- My thoughts regarding the internship

About the company

- Slice, which we call “the smart shopping assistant”, is a start-up revolutionizing the way digital commerce measured and specializing in package tracking, online spending history recording, and e-receipts extracting.

- Business Insider has called Slice one of their "Top 10 Productivity apps", and Inc.; magazine named Slice to their list of seven "Startups to Watch in 2012”.

- **How I found this internship opportunity?**

  -- Referred by my friend who is one of the data scientists at Slice Inc.

About my job

- I was working with a team of smart data analysts to support the sales and marketing by mining Slice's rich data set to reveal new insights about online transactions and shopping behavior. The majority of my task was to perform training data cleaning and data quality check as well as ad hoc tasks.

- I knew nothing about SQL and only knew some simple functions in EXCEL (e.g., sum, average functions, pivot table and pivot chart) before my interview. During the interview, I was asked to give self-assessment score of my skills on SQL and EXCEL.

- After two months' internship at Slice, I learned important skills in composing SQL string, operating functions in MS EXCEL.

Some major projects

- Used Excel to compare item descriptions from a text file with those from an excel file by using VLOOKUP function and returned the quantity of the items.
- Used SQL Server to extract 3700 items which have specific conditions to Excel and looked through these item titles to assign them into existed branches. During the assignment, I learned how to use nested "IF function" and "VLOOKUP function" to increase the efficiency and accuracy; I also created pivot tables for other data analysts for further use based on the result. I achieved the goal to reduce the percentage of "Other" category and cleaned the data.
- Used MySQL to help engineering team clean the training data set that has 172,720 rows. I wrote SQL queries to select items with relevant titles and then updated items whose category paths were incorrect with the correct paths. After three weeks of interactive work with the engineering team, I achieved a significant improvement in precision and recall indexes.

Precision and Recall

- In pattern recognition and information retrieval with binary classification:

	Precision	**Recall**
also called:	positive predictive value	sensitivity
	the fraction of retrieved instances that are relevant	the fraction of relevant instances that are retrieved
can be seen as:	a measure of exactness or quality	a measure of completeness or quantity
in simple terms:	high precision means that an algorithm returned substantially more relevant results than irrelevant	high recall means that an algorithm returned most of the relevant results

- Both precision and recall are therefore based on an understanding and measure of relevance.

Precision and Recall

- Machine learning?
- -> Statistical learning

Regression vs. Classification

	Regression	Classification
Y's (output):	numerical	categorical
a different way:	involves estimating or predicting a response	identifying group membership

- In statistics, if the null hypothesis is that all and only the relevant items are retrieved, absence of type I and type II errors corresponds respectively to maximum precision (no false positive) and maximum recall (no false negative).
- For more info, please refer to http://en.wikipedia.org/wiki/Precision_and_recall

My thoughts

Data quality issues:

- There might be duplicates in item titles but with different category paths.
- I might not cover all the training data set so that the data could not be totally cleaned.
- The "Other" category has a large proportion in the training data set.
- ...

My gains:

- After two months practical training, I learned to online Google things that I don't know, learned the power and conveniences of Excel functions, and learned how to write SQL queries with nested queries.
- In addition to the programming skills, I also practiced my communication skills, especially my spoken and listening English.

More about Slice Technologies, Inc.:

- LinkedIn: https://www.linkedin.com/company/project-slice
- App: download the Slice app to help you save time and save money!

Thank you for listening!