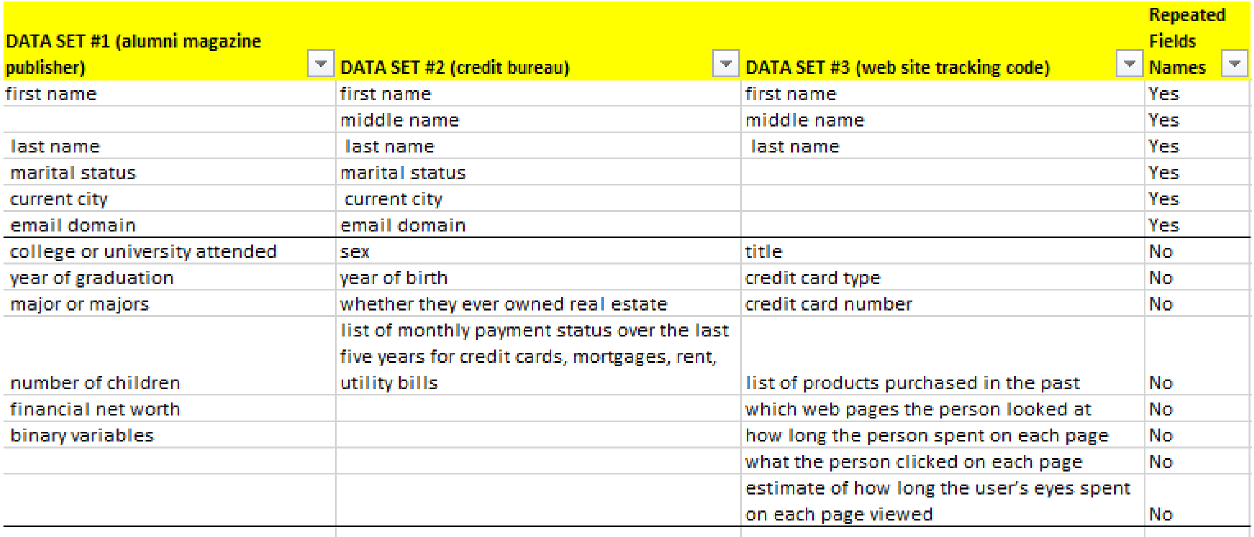
HW 10

Possible Recommendations

Dealing with Missing Data

Before we use any analytical models, it is vital to investigate the data fields in the three data sets. Some fields are repeated in either two or all the three data sets. For example, first name and last name is repeated in all the data sets, but middle name is repeated in only the second and the third data sets.



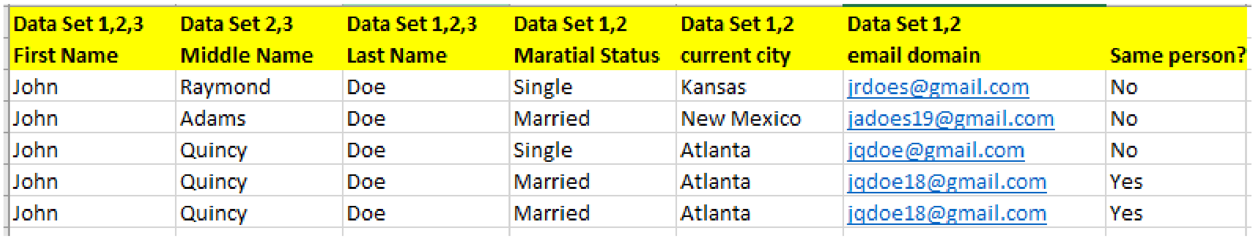
DATA SET #1 (alumni magazine publisher)	DATA SET #2 (credit bureau)	DATA SET #3 (web site tracking code)	Repeated Fields Names
first name	first name	first name	Yes
	middle name	middle name	Yes
last name	last name	last name	Yes
marital status	marital status		Yes
current city	current city		Yes
email domain	email domain		Yes
college or university attended	sex	title	No
year of graduation	year of birth	credit card type	No
major or majors	whether they ever owned real estate	credit card number	No
number of children	list of monthly payment status over the last five years for credit cards, mortgages, rent, utility bills	list of products purchased in the past	No
financial net worth		which web pages the person looked at	No
binary variables		how long the person spent on each page	No
		what the person clicked on each page	No
		estimate of how long the user's eyes spent on each page viewed	No

Identifying Middle Name for Data Set #1

Since middle name is missing in the first data set, one way of getting that information is to verify whether the first, middle (from second and third data sets), and last names match. For example, you can have an individual with the same first and last name but different middle name like John Raymond Doe, John Quincy Doe etc. Another issue we might come across is that there might be multiple people with the same first, middle, and last names. One way of solving this issue is to use email data. Since every person's email will be required to be unique, this will help us segregate amongst the people who have same first, middle, and last names. City and marital status data can also be used. This will enable us to get the actual middle name data at least for the people who have same identities.

Identifying marital status, current city, and email domain for Data Set #3

For the third data set, we compile a list of people who share the same first, middle, and last names from the three data sets and will have to use imputation to come up with an estimate for marital status, city, and email. This does rely heavily on intuition or guess work since we do not have other critical information such as email or last 4 digits of SSN. Since the third data set is related to ecommerce, the data owner will have to collect further information on shipping address and email. Having this information readily available will significantly aid in identifying people with same identities and will also reduce the time and effort spent on imputation.



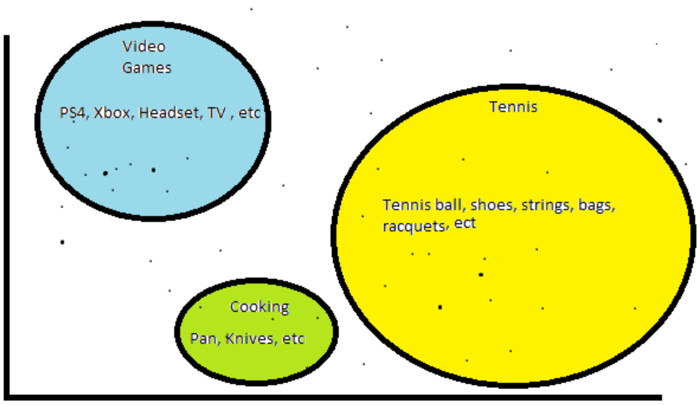
Data Set 1,2,3	Data Set 2,3	Data Set 1,2,3	Data Set 1,2	Data Set 1,2	Data Set 1,2	
First Name	Middle Name	Last Name	Maratial Status	current city	email domain	Same person?
John	Raymond	Doe	Single	Kansas	jrdoes@gmail.com	No
John	Adams	Doe	Married	New Mexico	jadoes19@gmail.com	No
John	Quincy	Doe	Single	Atlanta	jqdoe@gmail.com	No
John	Quincy	Doe	Married	Atlanta	jqdoe18@gmail.com	Yes
John	Quincy	Doe	Married	Atlanta	jqdoe18@gmail.com	Yes

***The picture above does not contain real data. This is just an example to show intuitively how to identify similar data.*

Combining Data Set to Create Value to the Organization

One way of combining the information from the first two data sets to is to identify the names of the people which shows up in both the data sets. College alumni association often partner with credit cards to make offers. Matching a person in the alumni magazine data set with a person in the credit bureau data set will help us in identifying the data required to use for an analytics model to determine what level of credit should be offered. With the list of monthly payment status and financial net worth, a support vector machine learning model can be used to classify high, medium, and low risk customers. Risk level indicate the probability of not paying credit card bills. Regression model can also be used to further identify the correlation between different variables. For example, if a person has attended multiple universities, does that mean that he/she will be carrying a lot of student loans which will in turn put them in them in high risk category. In the same way, we could also identify weather there is a correlation between the number of children and credit worthiness.

Another way to create value is to combine information from all three data sets. For users of the company's website, a list of the user's hobbies and interests taken from the alumni magazine data could suggest what types of products should be shown to the user. For this we can use clustering model to create clusters of activities. For example, tennis cluster can include products like tennis ball, tennis shorts, tennis band etc. The credit bureau data and past purchasing data could be used to determine what price level a product should be. For this a Regression model can be used. A user with better credit, higher monthly credit card expenses and more expensive purchases can be shown more expensive items and vice versa.


Video
Games
PS4, Xbox, Headset, TV , etc
Tennis
Tennis ball, shoes, strings, bags, racquets, ect
Cooking
Pan, Knives, etc

Using the Browsing Pattern Data to Create Value to the Organization

Design of experiments can be used to identify customer buying behavior. When a person first signs into an ecommerce website like amazon, the model could start out by showing pictures of a variety of different products. As it collected data about which ones the user was looking at more, the model could use a multi-armed bandit approach to recommend which products should cycle through the space for images. The model would trade off exploration, showing images of new products, and exploitation, showing images of products that are like ones the user looks at longer. If the user is just looking at two alternatives, A/B testing can also be used by displaying the two competing products and verifying which product gets the most clicks.