Alpha Analytics

Founded By:
Brian Walsh
Kevin Coppinger
Ryan Gill

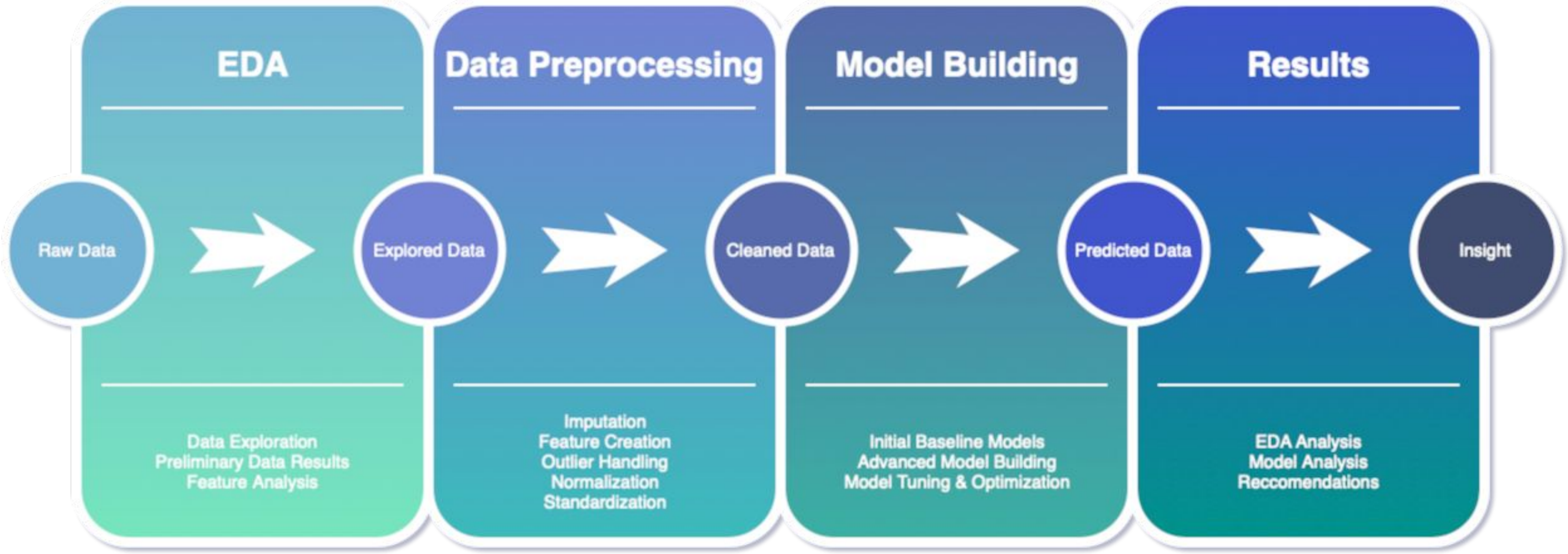
Setting the Stage

Goal of Our Campaign

The goal of our analytics campaign, per our customers request was:

To identify the patterns of behavior and demographic features that would allow us to identify a customer that was at a higher risk of defaulting on their payments.

Our Pipeline



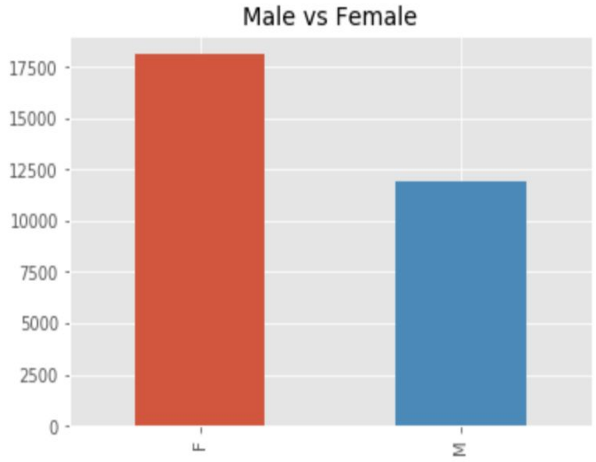
Preliminary Data Analysis

Data We Were Given

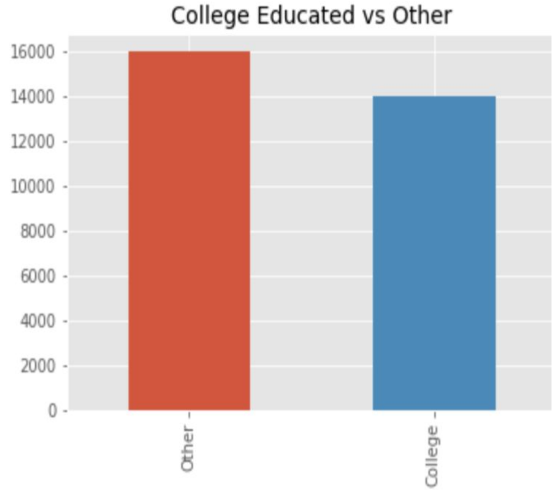
The data that the customer gave us was collected as a CSV file that included:

- 30,000 rows of client data
- 23 features/predictor variables
- 1 Target variable (default or not)
- 0 Missing Data

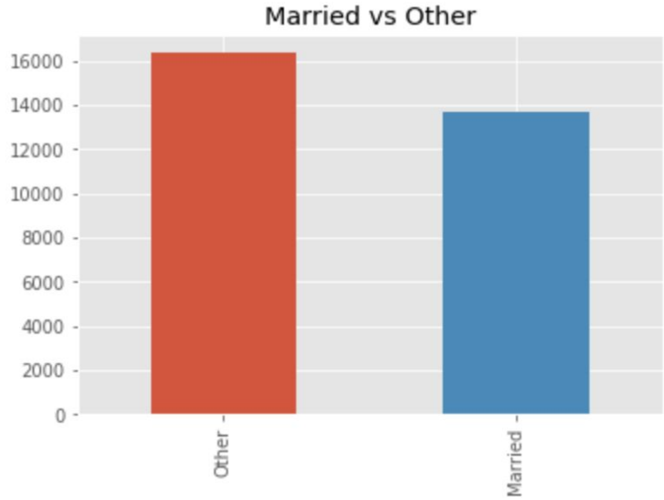
Preliminary Data Analysis



60% 40%

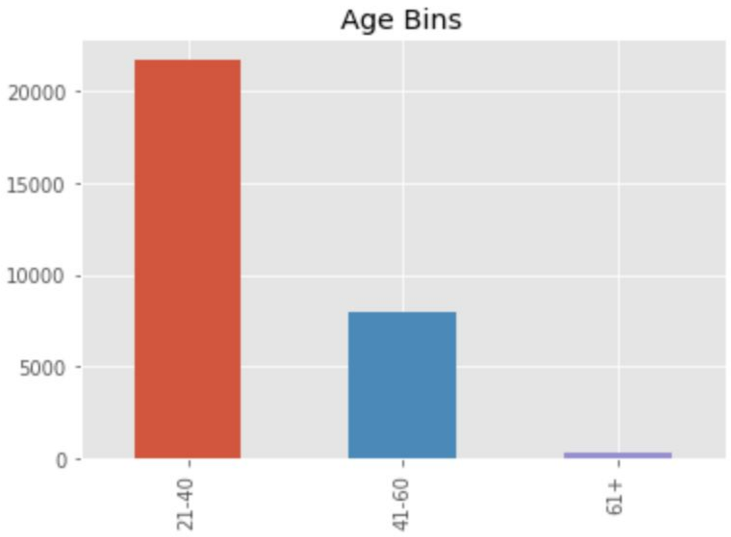


53% 47%

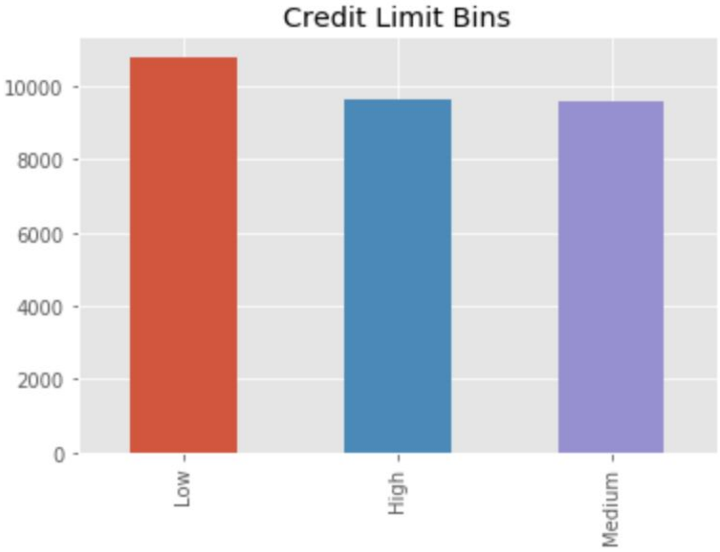


54% 46%

Preliminary Data Analysis Cont...

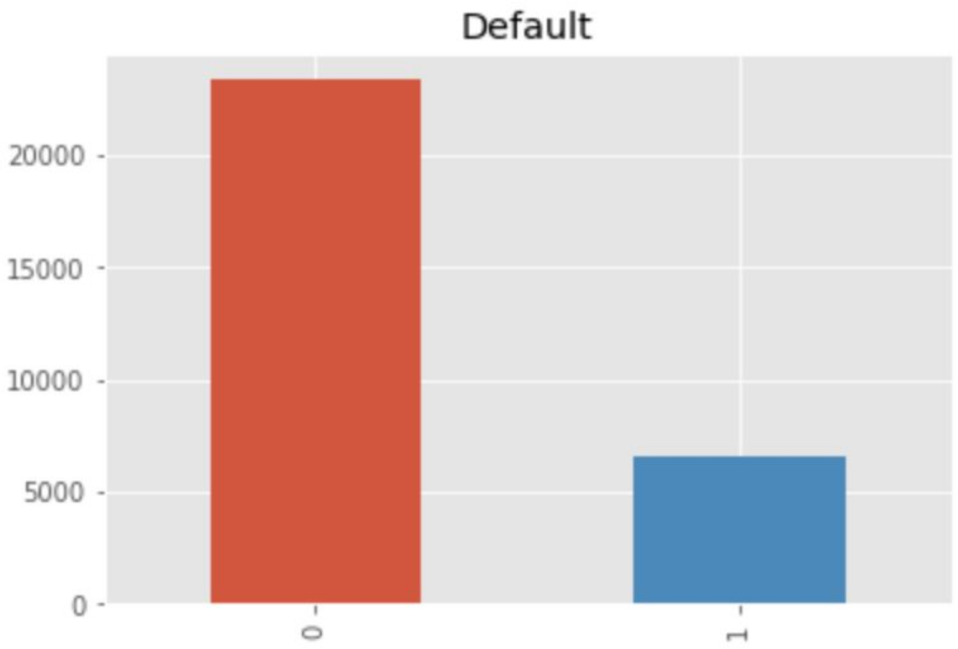


72% 27% 1%



36% 32% 32%

Preliminary Data Analysis (Defaults)

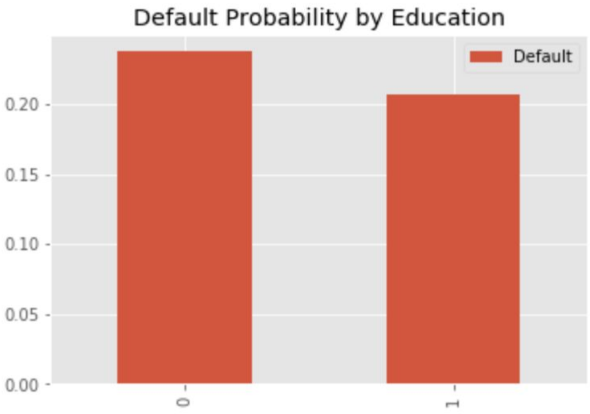


78.88%
Non-Default

22.12%
Default

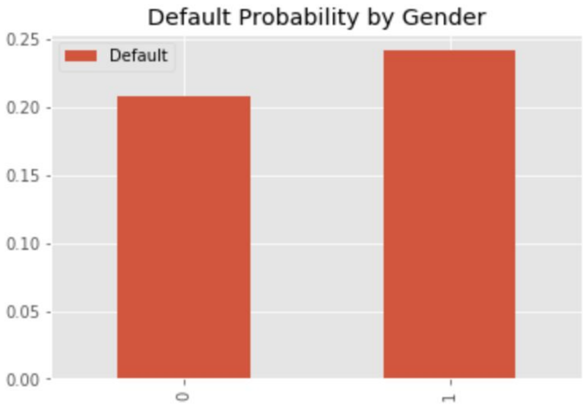
Preliminary Data Analysis (Defaults)

N = 6636 People



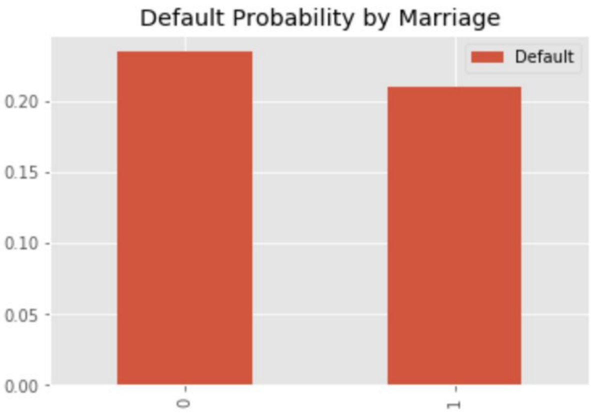
3,330
College

3,306
Other



3,763
Female

2,873
Male



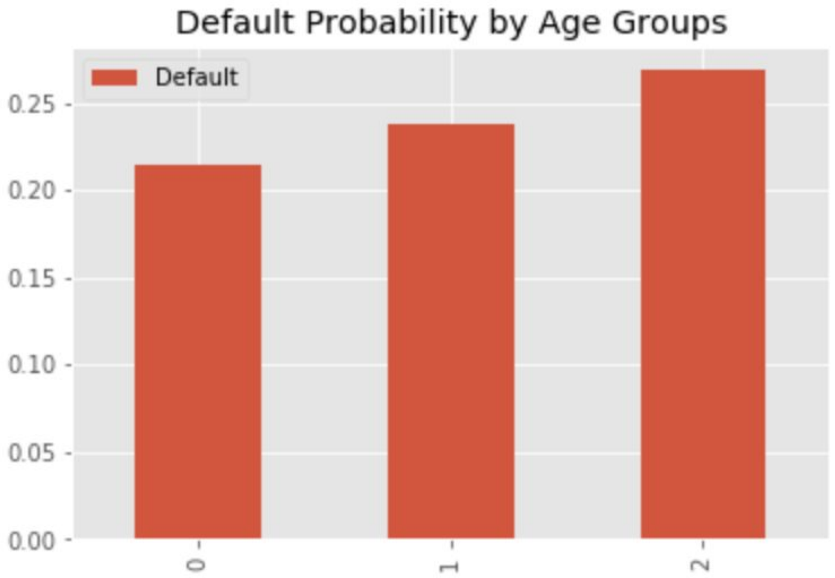
3,206
Married

3,430
Other

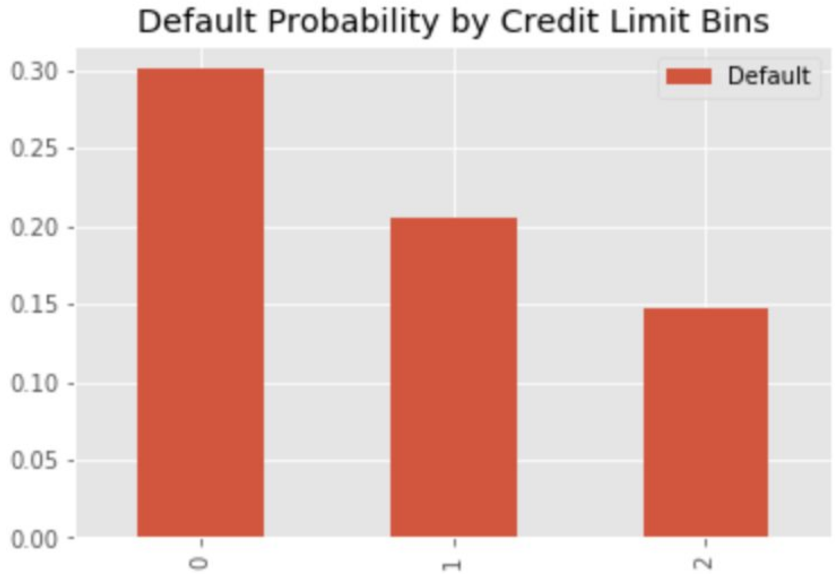
These are Conditional Probabilities

Preliminary Data Analysis (Defaults)

N = 6636 People



4,660 21-40 | 1,903 41-60 | 73 61+



3,246 Low | 1,973 High | 1,417 Medium

These are Conditional Probabilities

Data Pre-Processing

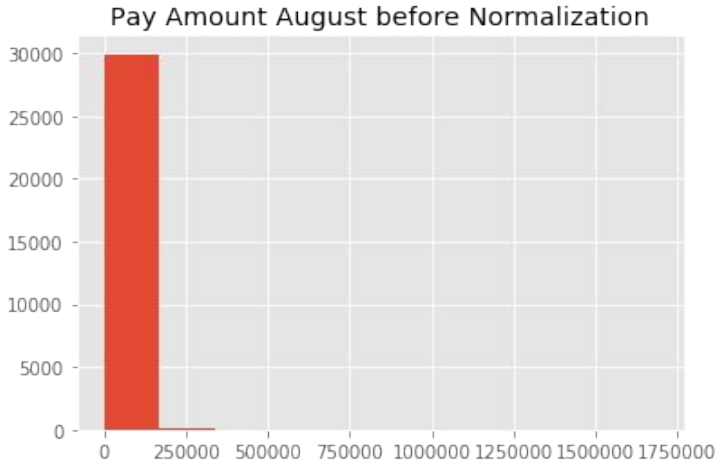
Data Preprocessing - Imputation, Feature Creation and Manipulation

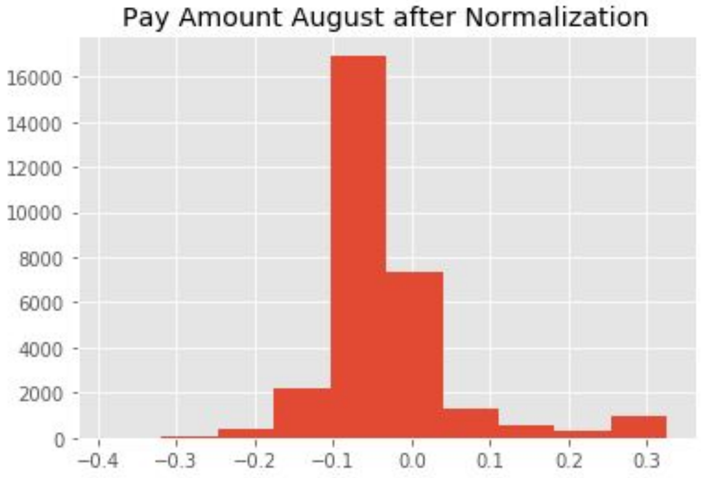
- No Missing Values: No Imputation Needed
- Created a Target and Predictor DF
    - Segregated out default column
- Had to oversample our target variable
- Binned age, education, gender, marital status and credit limit
- Created columns for credit utilization
    - Percentage of bill amount and credit limit
- Created column for total pay amount
- Created column for total bill amount
- Created columns for going over the credit limit

Data Preprocessing - Normalization

- Used histograms for each feature to check for skewness
- Then checked for the skewness of each feature in numerical form
  - Pay Amounts heavily skewed
- Used Sklearn's normalize function to remove the skewness
  - Pay Amounts still skewed
- Used a log transformation because the Pay Amounts were right skewed

Skewed Histogram





Data Preprocessing - Standardization

- Data must be normalized before it can be standardized
- Used minmaxscaler to have all features values be between 0 and 1
- Done to have all features on the same scale

Feature Selection

- Want to reduce the dimensionality of the data
- Used RFE and a correlation analysis
- Compared the results to select the most important features
- Used PCA on the correlation variables because it gave us the best results

RFE vs. Correlation Features

RFE	Correlation
Credit Limit	**Pay Sept**
Pay Sept	**Pay Aug**
Pay Aug	Pay Jul
Bill Amount Sept	Pay Jun
Bill Amount Aug	Pay May
Bill Amount Jul	Pay Apr
Bill Amount May	Pay Amount Sept
Pay Amount June	

Initial Baseline Models

- Created logistic regression models using our initial dataframe, RFE variables, Correlation variables and PCA variables
- Wanted to get a baseline to go off of for our future models
- Baselines were around 80% accurate but were most likely overfitting

Advanced Modeling

Advanced Model Building - Ensemble Bagging

- Used Bagging with Decision Tree Classifier
- Initially overfit the model ~ 94% cross validation score
- Defined max_depth/num_trees to reduce overfitting
    - Issue was our trees were too deep and had too many
- Reduced cross validation score to around 73%

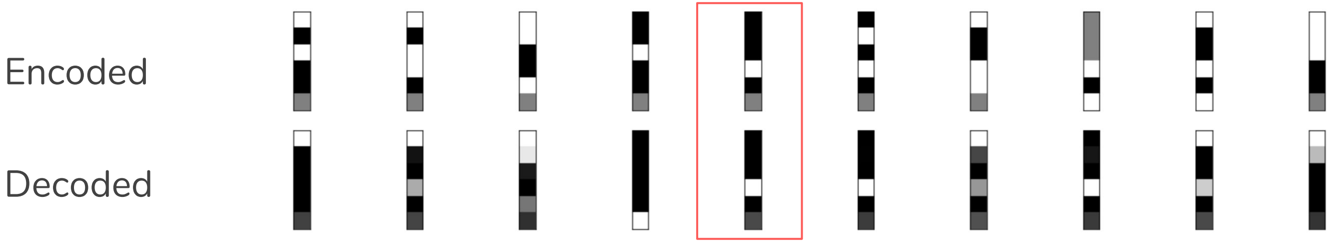
Advanced Model Building - Artificial Neural Net

- Very Good Model
- Used Oversampling on Target Variable
- Used Keras Deep Learning Library to build ANN
    - Built on Tensorflow Backend

	Precision	Recall	F1-Score
Not Default	.88	.80	.84
Default	.46	.61	.52
Avg	.79	.76	.77

Advanced Model Building - Auto Encoder

- Attempted to build a time series model using AE
- Looked at debt accumulation over time as a trend
- Could not get model to work because output was a continuous variable, not categorical



Advance Model Building - XGBoost

- Best Model
- Execution speed is high compared to other models
- Model performance is usually better than other models
- Created Baseline Model (76% Accuracy)
- Used Random Search to optimize 4 different hyperparameters

Advance Model Building - XGBoost Results

Model	Output	Precision	Recall	F1	Accuracy	CV Accuracy
Original	0	.88	.82	.85	.77	.83
Original	1	.48	**.61**	**.54**		
	Avg	**.80**	.77	.78		
Optimized	0	.88	.83	.85	.78	.82
Optimized	1	.48	.57	.52		
	Avg	.79	**.78**	.78		

Results and Analysis

- Based on our model, we identified the top 7 factors impacting the model

Feature	Importance
Bill Amount September*	7.5%
Credit Limit*	7.0%
Outstanding Debt	6.1%
Bill Amount August*	5.9%
Pay Amount August	5.9%
Bill Amount April	5.7%
Pay Amount September	5.6%

Identifying New Customers Default Risk

- Age, Credit Limit, and Education are the biggest factors impacting Default or Not
    - Highest Risk Groups include:
        - Those over the Age of 61
        - Those with a Low Credit Limit
        - Those who are College educated
- Overall Trends Indicate that as you get older, the likelihood of defaulting increases
- 2x as likely to default in the ‘low credit’ range than ‘medium credit’ range
- Trends also suggest that the higher your credit limit, the less likely you default
    - Our data does not suggest that the older you are, the higher your credit limit becomes

Identifying Current Customers Default Risks

- Look at accumulation of debt over time (just the trend)
  - If debt is increasing MoM than more likely to default
- Biggest factor is their Bill Amount the Previous Month
  - As Indicated by Bill Amt September
- Credit Limit and Overall Outstanding Debt is important factor as well
- Their payment types were not indicative to defaulting in our model

Questions?