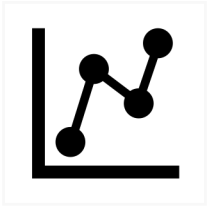
PREDICTING LOAN LIKELIHOOD

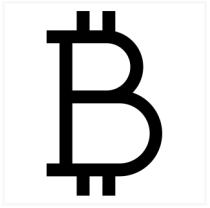
SHYAM MOHAN KIZHAKEKARA
DATA ANALYST/DATA MODELLER
SHYAMMOHAN.KIZHAKEKARA@BOI.COM

FORMULATING THE BUSINESS PROBLEM

Can the data help improve the business and customer experience?



Can the data provide key business insights to strengthen marketing strategies?



Can the insights be used to invest wisely and maximize the return on investment?

FORMULATING THE TECHNICAL PROBLEM

How can the key business insights be derived from the data?

How can the accuracy of these insights be improved?

How can the whole idea be designed and integrated as a continuous process?

CUSTOMER DATA SNAPSHOT

A	B	C	D	E	F	G	H	I	J	K	L
ClientID	Age	Gender	County	LoanHeldBefore	NoOfProductsHeld	AvgTxnAmt	NoOfTxns	LastTxnAmt	LoanFlag	IncomeCategory	MerCategory
1	36	1	Cork	1	4	58	0		0	Lower Middle	Unknown
2	43	1	Cavan	0	4	2663	17	83.66	0	Low	Professional Services
3	32	0	Dublin	0	2	46	25	526.18	0	Lower Middle	Accommodation
4	52	1	Louth	1	2	0	13	70.68	0	Upper	Hardware
5	63	0	Kilkenny	0	1	126	39	259.07	0	Upper Middle	Retail
6	33	0	Louth	0	2	338	39	257.15	0	Lower Middle	Accommodation
7	25	0	Dublin	0	1	0	16	647.02	0	Lower Middle	Transport
8	68	1	Kildare	0	4	977	17	605.96	0	Upper Middle	Transport
9	24	0	Cork	0	4	22	5	45	0	Upper Middle	Retail
10	34	0	Dublin	0	2	115	5	330.13	0	Lower Middle	Transport
11	54	0	Dublin	0	4	91	0		0	Lower Middle	Unknown
12	64	0	Kilkenny	0	2	0	0		0	Lower Middle	Unknown
13	31	1	Carlow	1	1	232	0		0	Upper Middle	Unknown
14	33	1	Dublin	0	2	0	23	348.19	0	Upper Middle	Health
15	37	0	Offaly	0	5	72	0		0	Lower Middle	Unknown
16	29	0	Dublin	0	2	392	0		0	Upper Middle	Unknown
17	51	1	Dublin	0	2	0	11	375.48	0	Lower Middle	Accommodation
18	23	1	Cavan	0	3	0	0		0	Lower Middle	Unknown
19	30	0	Carlow	1	2	324	0		0	Lower Middle	Unknown
20	27	0	Galway	0	1	95	37	330.05	0	Upper Middle	Transport
21	48	1	Dublin	1	3	669	21	694.31	0	Lower Middle	Retail
22	48	0	Wicklow	1	4	0	17	464.99	0	Lower Middle	Retail
23	41	1	Dublin	0	2	0	63	490.03	0	Lower Middle	Transport
24	57	1	Cork	0	2	0	11	890	0	Lower Middle	Office Equipment
25	35	0	Dublin	0	1	1159	35	137.42	1	Lower Middle	Professional Services
26	50	1	Dublin	0	5	0	55	626.88	0	Lower Middle	Accommodation
27	52	0	Offaly	0	0	0	19	339.79	0	Upper Middle	Professional Services
28	46	0	Dublin	-1	1	280	20	450.07	0	Lower Middle	Retail
29	22	1	Cavan	0	2	1368	30	314.89	0	Low	Accommodation
30	45	0	Dublin	0	2	0	15	125.05	0	Lower Middle	Misc
31	21	1	Dublin	0	3	4	8	82.78	0	Lower Middle	Professional Services
32	23	1	Kilkenny	0	1	350	26	872.78	0	Lower Middle	Professional Services

Above is a snapshot of our final dataset compiled by merging the different CSV files in scope. It has also undergone the first level of pre-processing.

UNDERSTANDING THE CUSTOMER DATA

ClientID – Unique identifier for a Customer.

Age – Age of the Customer as per the records

Gender – Recorded Gender of the Customer

County – The County in which the Customer resides in.

LoanHeldBefore – Depicts whether the customer has held a loan with the Bank.

NoOfProductsHeld – No. of products provided by the Bank, that is used by the customer.

AvgTxnAmt – Average transaction amount of the customer.

NoOfTxns – No. of transactions involving or initiated by the Customer.

UNDERSTANDING THE CUSTOMER DATA

LastTxnAmt – The last recorded transaction amount of the Customer.

IncomeCategory – The slab to which the Customer is classified based on his annual income.

MerCategory– The services category assigned to a merchant code

LoanFlag – The indicator that depicts whether the Customer has availed a loan or not.

KEY ASSUMPTIONS

Income Category

The income slabs have been labelled as follows.

0 – 10,000:	10,001 – 40,000:	40,001 – 60,000:	60,001 – 100,000:	100,001 & Above:
Low	Lower Middle	Upper	Upper Middle	High

KEY ASSUMPTIONS

Likelihood Categories

The likelihood categories are defined as follows.

85% – 100%:	70% – 85%:	50% – 70%:	25% – 50%:	0% – 25%:
Very High	High	Medium	Low	Very Low

THE DATA PIPELINE

Data Ingestion and Data Pre-Processing

Find a set of features in the data that contribute the most to the prediction variable.

Build a classification model that predicts which customers are more likely to take a loan.

Calculate the loan uptake percentage to classify the customers into the likelihood-categories.

FINDING A FEW ANSWERS

1. How Many Customers above 50 years old have taken up a loan?

98

2. How Many Females aged 30 to 40 have more than 2 products?

622

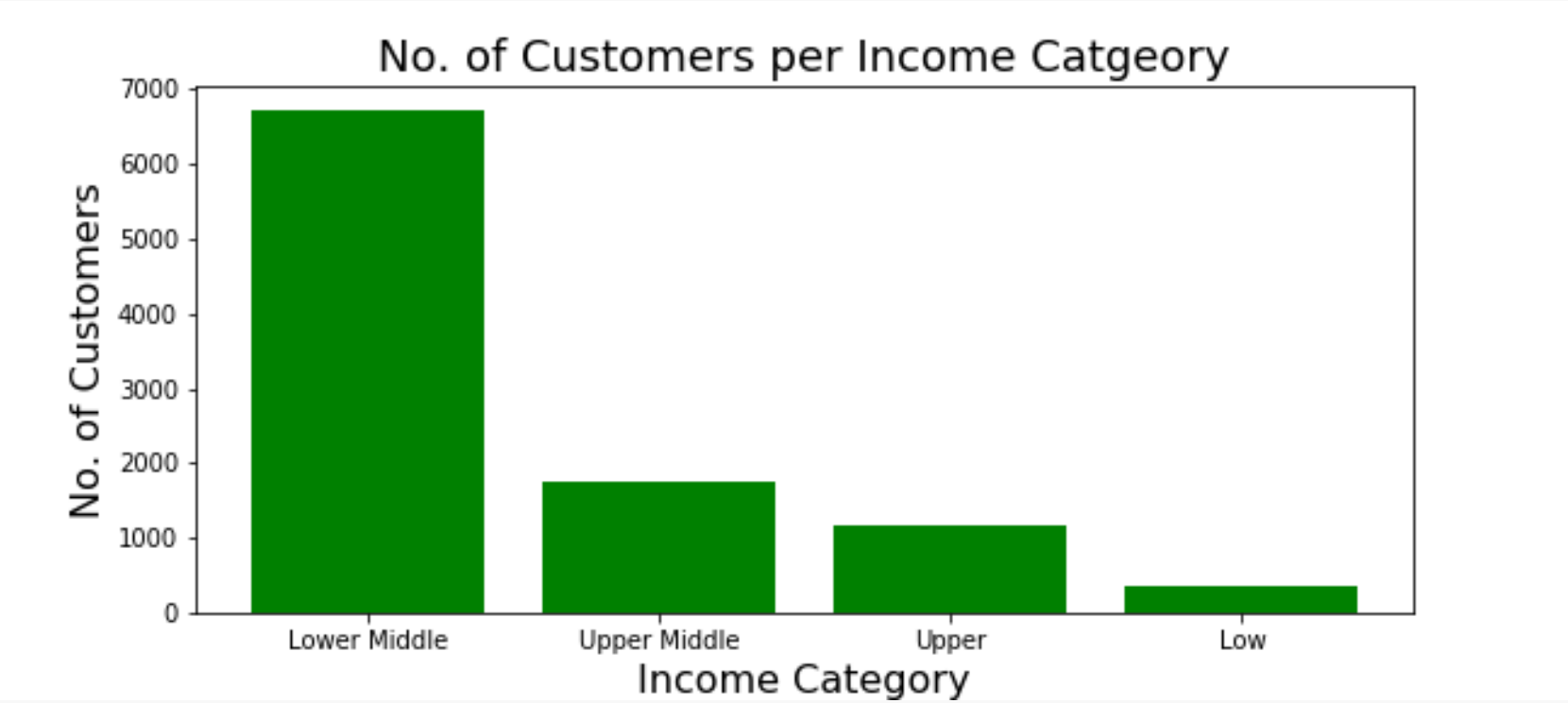
3. What is the average number of Current Account(CA) Transactions for males who had a previous Loans?

19 [Actual Value: 18.646]

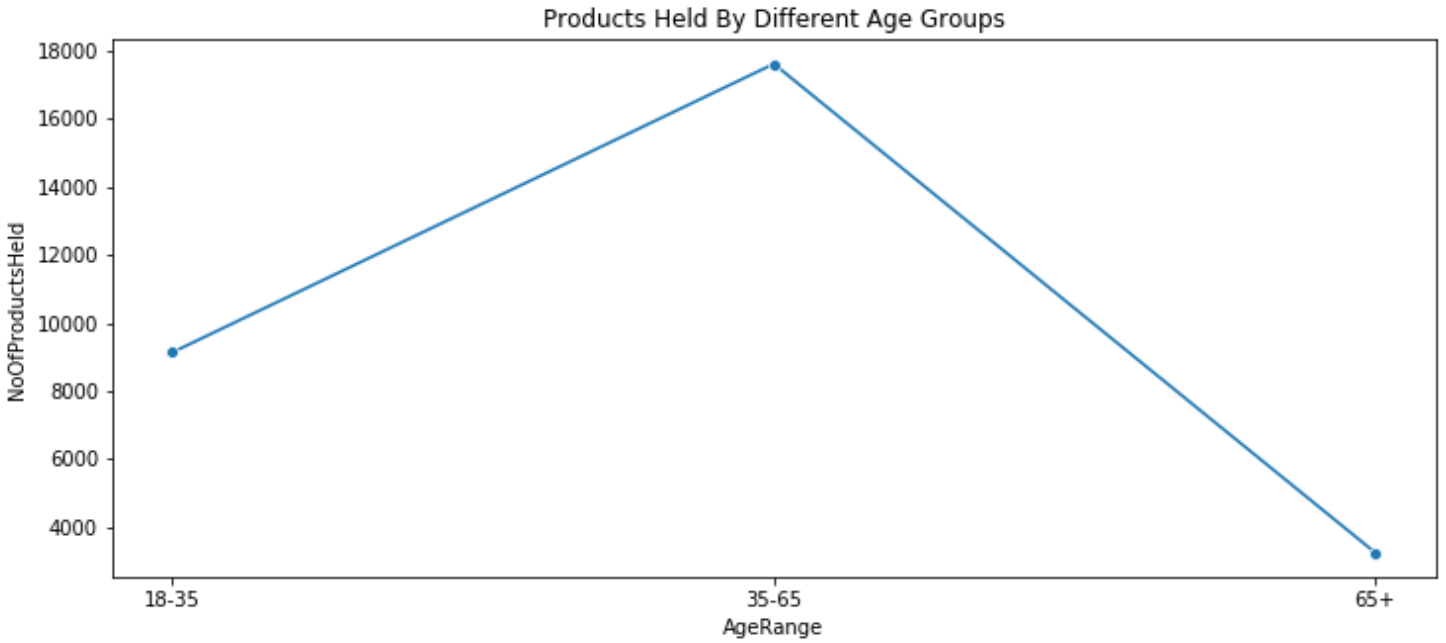
4. How many females did not have a previous loans and who are aged

- Less than 20 - 0
- 21 to 40 -1534
- 40+ - 2003

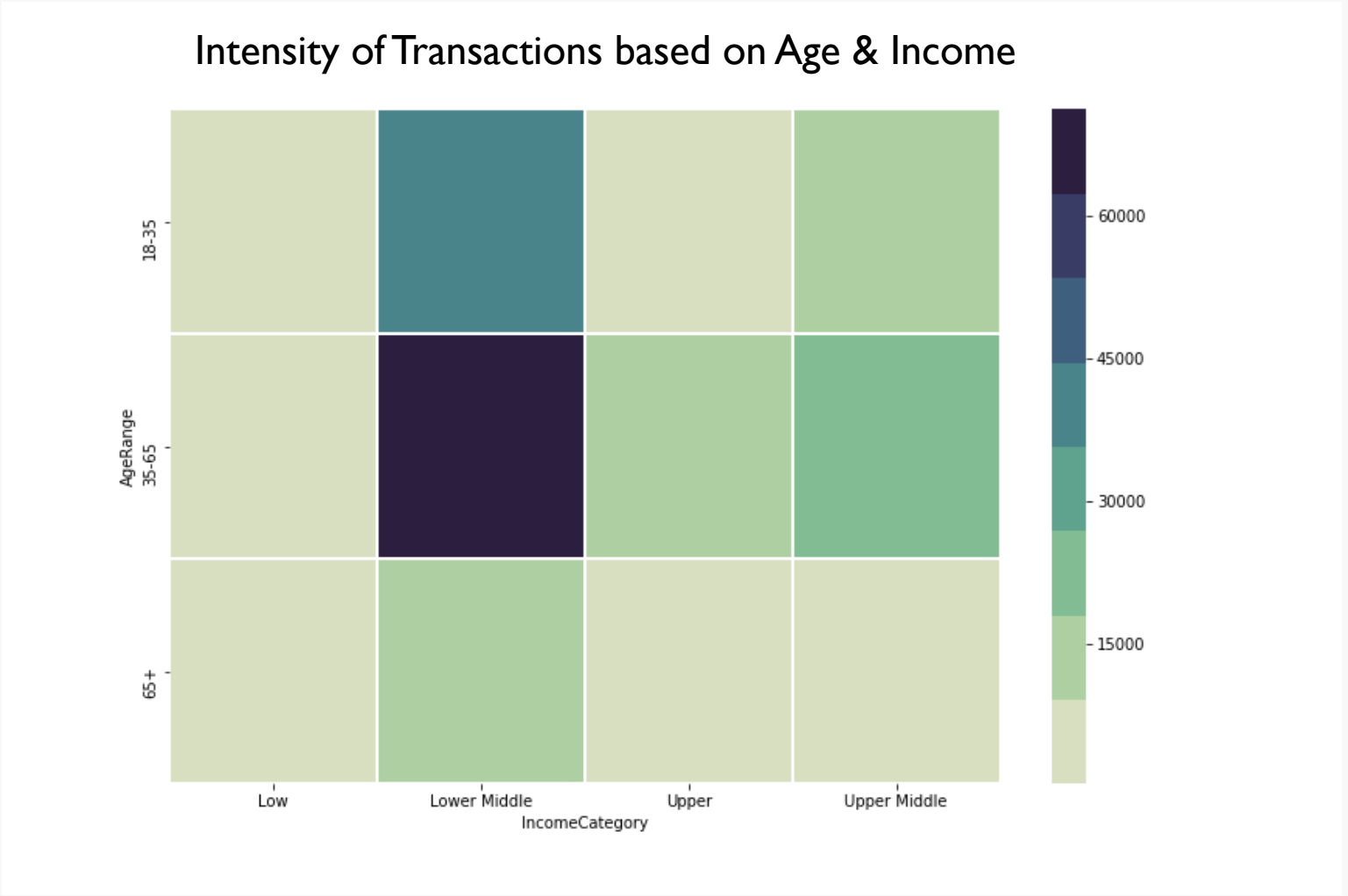
SOME KEY POINTERS FROM THE DATA



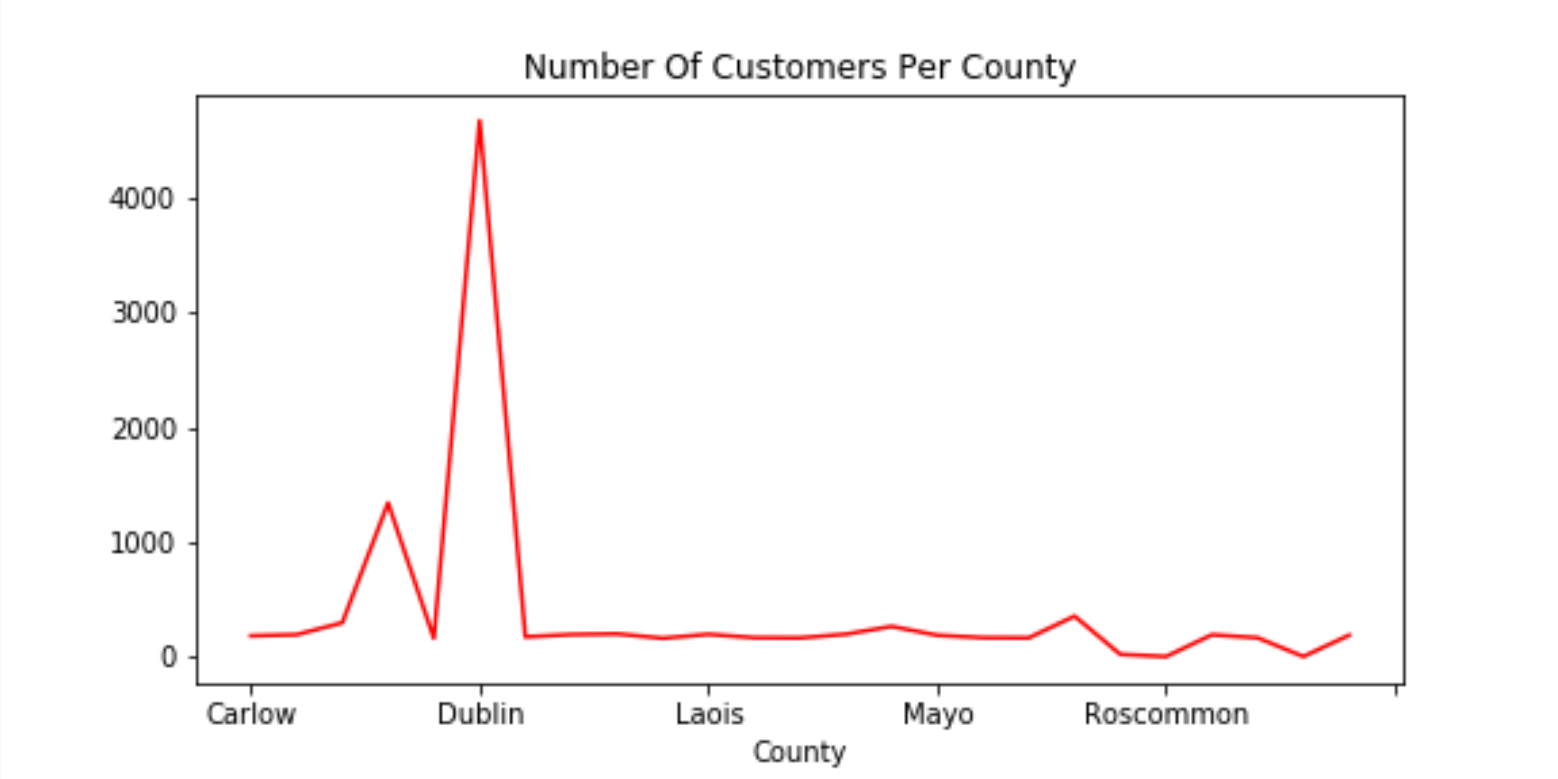
SOME KEY POINTERS FROM THE DATA



SOME KEY POINTERS FROM THE DATA



SOME KEY POINTERS FROM THE DATA



HAD THERE BEEN MORE DATA…

- **Transactions from the past 12 Months for each Customer**
  - Date of Transactions
    - A time-series visualization could give us the specific months in which a customer spends more.
  - Transaction Amounts
    - Idea about a Customer's expenditure per day, per month etc. could be retrieved.
- **Details of Existing Loan**
  - Existing Loan Parameters
    - Details of existing loans like Outstanding Balance Amount, Defaults etc., would have helped identify more patterns.

BUILDING OUR MODEL

Data Pre-Processing

- Drop irrelevant rows and columns
- Resolve Class Imbalance
- Data Imputation & Normalization
- Handle Categorical Variables

Feature Selection

- Selecting the best set of features in the data, that contribute most to the prediction of target variable (i.e., the loan likelihood).

BUILDING OUR MODEL

Performance Tuning

- From the list of values for available parameters of the algorithm (Logistic Regression, in our case), find the best parameter-values that gives us a good accuracy score. This is done using GridSearchCV.

Model Evaluation

- Use techniques like cross validation to improve the generalizability of the model and evaluate the predictive performance of the model by comparing the predicted values and actual values, and check for any bias.

GOODNESS OF THE MODEL

The model predicted the loan likelihoods with an accuracy rate of approximately 96.185%.

The loan uptake rates were calculated using the log-odds predicted by the Logistic Regression model, on the best feature subset provided by Random Forest Classifier.

Based on the uptake rate values, appropriate likelihood categories were assigned to each customer.

KEY INSIGHTS

- After assigning the likelihood category to each customer based on the propensity score, we could summarise the results as shown below.

```
df_loan_likelihoods['Likelihood Category'].value_counts()
```

```
Very Low Likelihood     1590
Low Likelihood           155
Very High Likelihood     113
Medium Likelihood         92
High Likelihood           50
Name: Likelihood Category, dtype: int64
```

- A fair share of customers are in the category 'Very High Likelihood' and 'High Likelihood', followed by a set of 'Medium Likelihood' ones.
- The business needn't spend time and money to target the customers in 'Very High Likelihood' category.

NEXT STEPS

The primary objective should be to target the customers in the category 'High Likelihood', through the Bank's marketing campaigns.

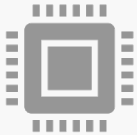
Considerably good offers can push those customers into the 'Very High Likelihood' category and would eventually benefit the Bank.

Tailored campaigns focusing on the customers in 'Medium Likelihood' categories must be launched.

Awareness programmes must be launched for the customers in 'Low Likelihood' and 'Very Low Likelihood' categories, to make them more understand the Bank's products and services.

FUTURE WORK

Classify the customers into different groups and define the best suitable marketing strategies for each group.



Build robust recommender systems for each group based on their web-logs (obtained through browser cookies).

Construct a data pipeline involving all the relevant processes and deploy it as a continuous process for any incoming customer records.

Continuously improve the model and its performance, as more and more data becomes available for training.

QUESTIONS

"If you torture data long enough, it will eventually confess to anything."

– Ronald Coase

"If you don't reveal some insights soon, I'm going to be forced to slice, dice, and drill!"