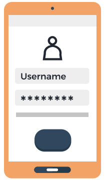
LHL_mini_project_V

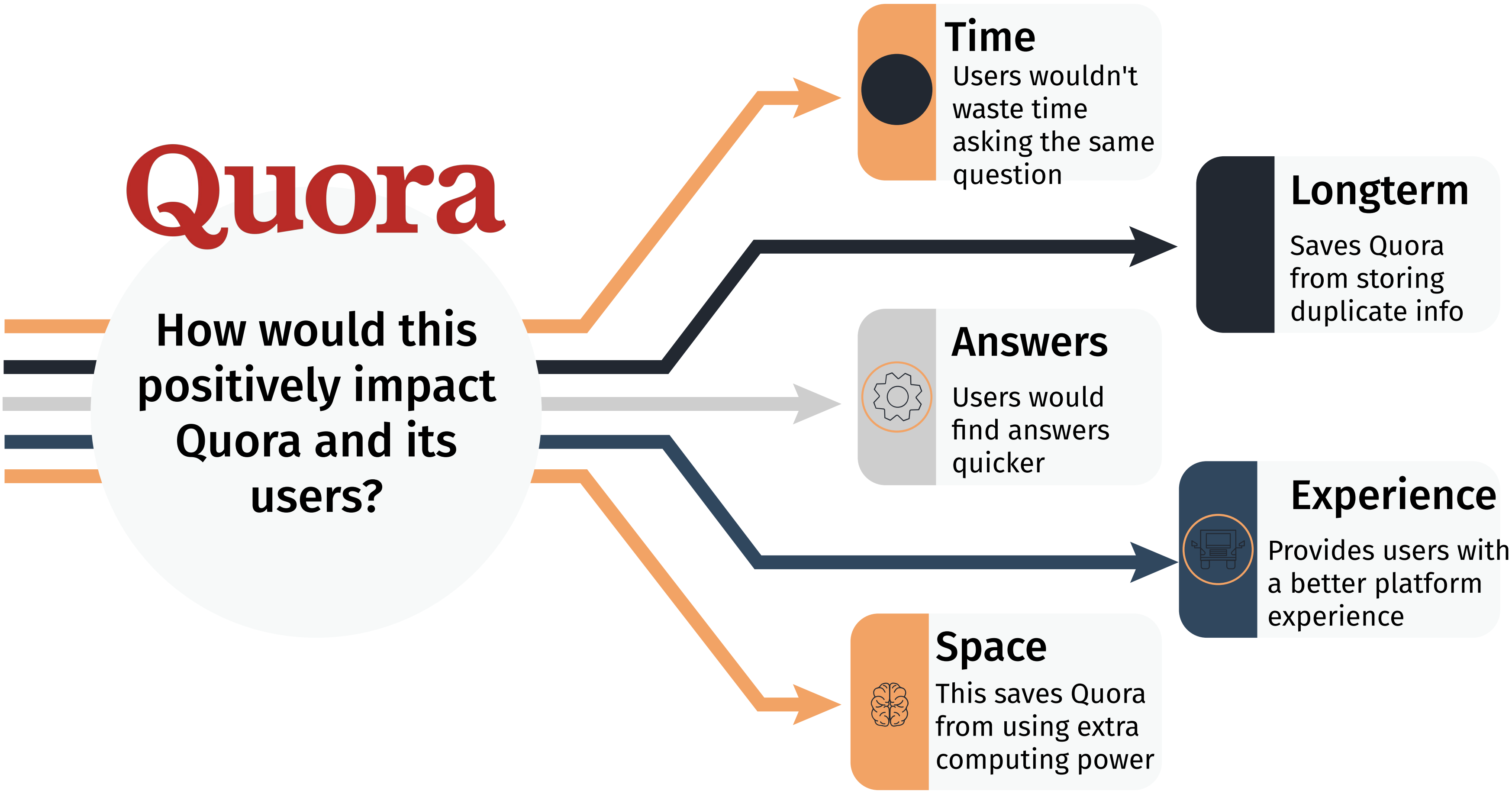
By: Jesse Randolph



Linktree:
https://linktr.ee/jesserndlph

Project Goal: Detecting duplicate Questions on Quora

Over 3 million people visit Quora every day, so it's inevitable that many people ask similar (or the same) questions.



Quora

How would this positively impact Quora and its users?

Time

Users wouldn't waste time asking the same question

Longterm

Saves Quora from storing duplicate info

Answers

Users would find answers quicker

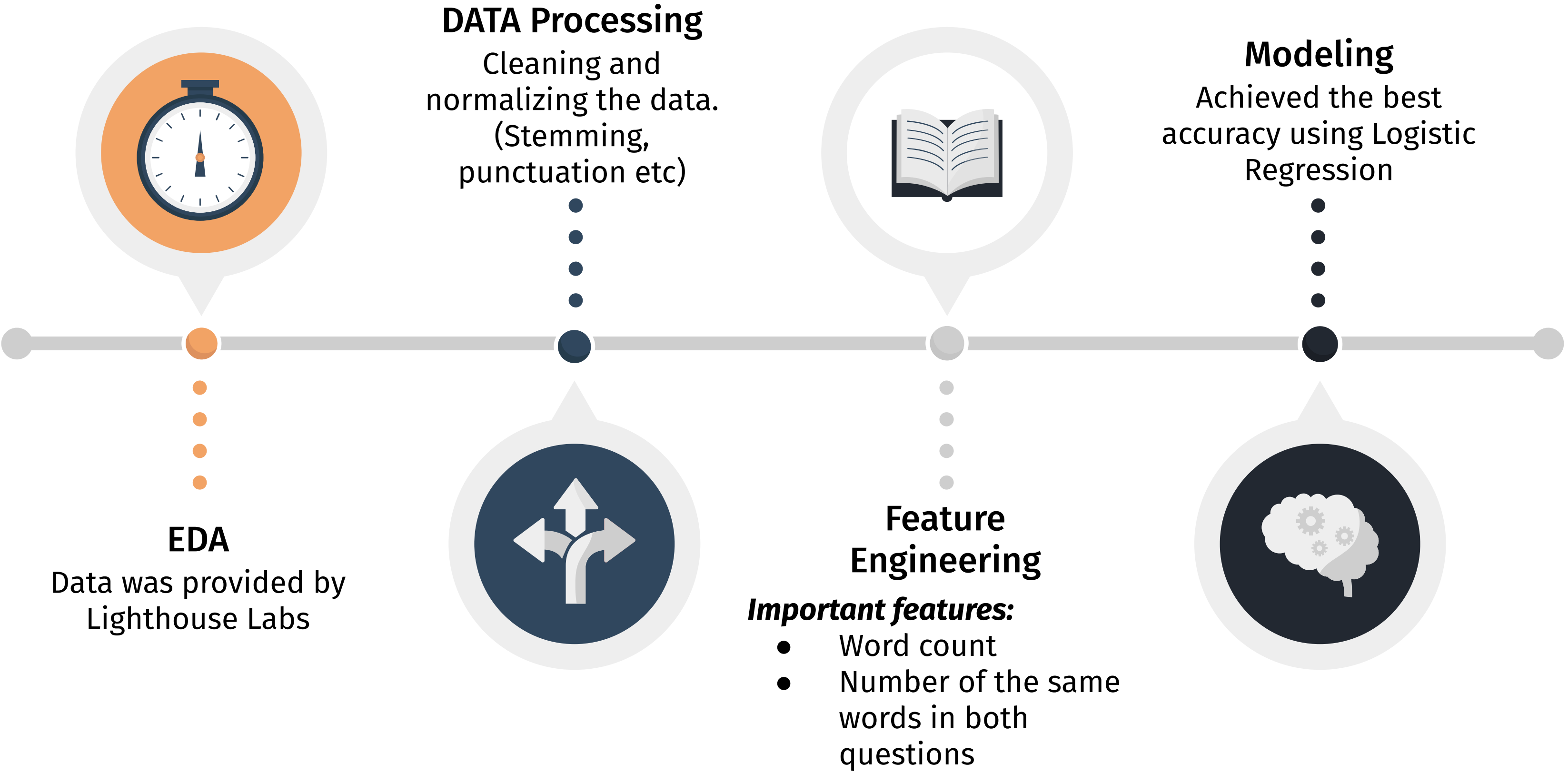
Experience

Provides users with a better platform experience

Space

This saves Quora from using extra computing power

Project Roadmap



EDA

Data was provided by Lighthouse Labs

DATA Processing

Cleaning and normalizing the data. (Stemming, punctuation etc)

Feature Engineering

Important features:

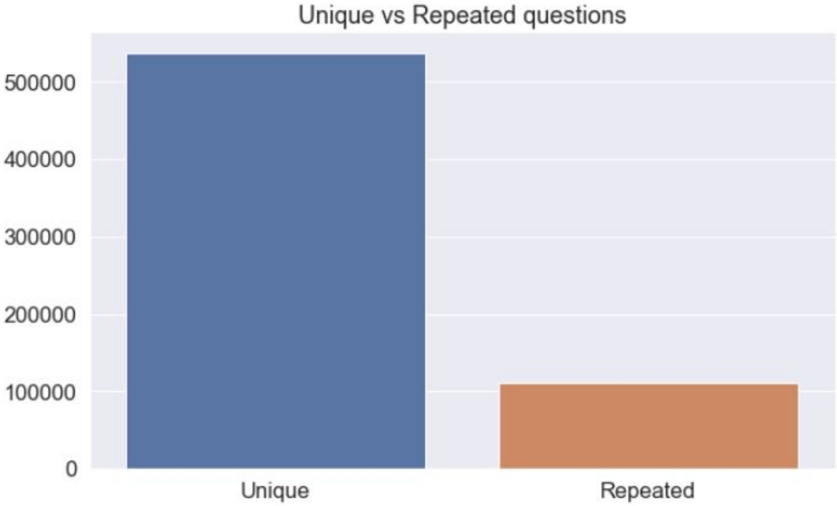
- Word count
- Number of the same words in both questions

Modeling

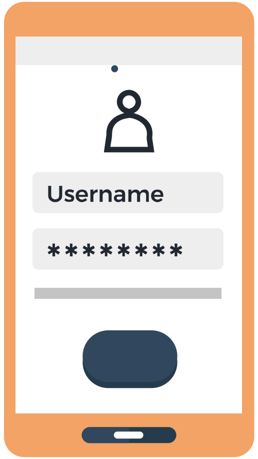
Achieved the best accuracy using Logistic Regression

Exploratory DATA Analysis

```
Total number of  Unique Questions are:
537933
Number of unique questions that appear more
than one time: 111780 (20.77953945937505%)
Max number of times a single question is
repeated: 157
```



DATA Preparation



Methods:

- Normalizing
- Stop Words
- Removing Punctuation
- Tokenization

From:

“Would you ever date Ryan Gosling’s left Toe?”

To:
would date ryan gosling left toe

Feature Engineering

Features models were Trained on:

```
features = df3[['is_duplicate',
'q1_word_num',
'q2_word_num',
'total_word_num',
'differ_word_num',
'total_q1q2_unique_word_num',
'total_nostop_len_diff']]
```

Model: sklearn Logistic Regression

73%

This model had the best accuracy score out of all of the algorithms used

Thanks for Listening!

Questions?

LHL_mini_project_V

By: Jesse Randolph

Linktree:
https://linktr.ee/jesserndlph

**Graphics by Slidesgo*