Statistics Round

Duration: 40 Minutes
Difficulty: Medium
Domain: Analytics

Problem

Facebook runs a games platform that allows users to play games with friends. The platform includes popular games such as Words with Friends, Candy Crush and Tetris. The business team wants to understand revenue drivers on games monetized via in-app purchases. Address the following statistical questions to assist the business team:

#1 - How would you use the revenue data to understand key factors driving the overall revenue growth on Facebook's games?

Title	Year	Country	Payment Method	Downloads	Revenue
Tetris	2015	S. Korea	PayPal	4M	$4M
Tetris	2015	US	PayPal	10M	$6M
Tetris	2016	S. Korea	PayPal	5M	$3M
Tetris	2016	US	PayPal	6M	$5M
Candy Crush	2015	S. Korea	PayPal	10M	$10M
Candy Crush	2015	US	PayPal	12M	$15M
….	….	….	….	….	….

#2 - Suppose that a linear model was fitted on categorical variables, resulting in 0.99 R-squared. Should you use this model?

Solution

> **#1 -** How would you use the revenue data to understand key factors driving the overall revenue growth on Facebook's games?

[Candidate] To ideate a solution, I'd like to ask questions to clarify assumptions. First, is the business goal to address factors driving revenue changes?

[Interviewer] Yes. Given the data you have, you will need to frame key business questions that can be addressed. Here's an example: *Games revenue on the platform increased by 20% from 2014 to 2016 because of A, B, C factors.*

[Candidate] Thank you for the clarification. I have another question. Can I assume the following for each variable?

1. The Title column contains millions of unique game titles.
2. The Year column contains years 2014, 2015, and 2016.
3. The Country column contains 180+ countries
4. The Payment Method contains payment methods applied to make in-app purchases and the values are PayPal, credit and debit.
5. The Downloads and Revenues are self-explanatory.

[Interviewer] Yes, all of those are valid assumptions about your data.

[Candidate] I'd like to begin with a simple analysis that does not entail a model. To understand top drivers that impact overall revenues, I can apply segmentations based on granularity and time comparison.

For instance, suppose we want to understand top revenue drivers by countries from the year 2015 to 2016. I can aggregate the data by country to create the following table:

Country	Baseline Revenue	YoY Pct Change
S. Korea	$10M	20%
US	$16M	15%
Canada	$3M	18%
Spain	$0.5M	-20%
….	….	….

[Interviewer] Can you elaborate on what you mean by “Baseline Revenue” and “YoY Pct Change?”

[Candidate] Certainly. The Baseline Revenue is the total revenue generated across titles by country in 2015. The YoY Pct Change is the % change in the total revenue by country from 2015 to 2016.

With this transformed data, I can now rank the country based on top drivers. For instance, I can take the top percentiles on both variables to get a list of top countries that contributed largely to the total revenues of the Facebook’s games platform.

[Interviewer] What if your granularity is more than just a country? Suppose you want to understand revenue drivers by country and payment method. How would you approach the analysis then?

[Candidate] Once again, the sample analysis can be applied. First, create the following data below:

Country	Payment Method	Baseline Revenue	YoY Pct Change
S. Korea	PayPal	$6M	15%
US	PayPal	$3M	5%
Canada	PayPal	$1M	3%
Spain	PayPal	$0.5M	-14%
S. Korea	Credit	$20M	8%
US	Credit	$12M	10%
Canada	Credit	$1M	14%
Spain	Credit	$0.2M	-2%
….	….	….	….

Then, apply the same ranking using percentiles to understand the drivers with the most and least contributions to the overall revenue changes from 2015 to 2016. For instance, for each of the indices, compute quantile distribution to create three buckets - High (<= 33th percentile), Medium (34th percentile <= x < 67th percentile), low (>= 67th percentile).

This will create a total of six groupings:

Buckets	Baseline Revenue	YoY Pct Change	Country-Payments
1	High	High	US-Credit, Ireland-Credit, UK-PayPal …...
2	High	Medium	…..
3	High	Low	…..
4	Low	High	…..
5	Low	Medium	…..
6	Low	Low	…..

[Interviewer] How would you report this analysis to the business team?

[Candidate] Based on bucketing, I can address key business questions such as *which drivers (country-payment pairs) contributed the largest in revenue growth on Facebook games from 2015 to 2016*? *Drivers such as US-Credit and Ireland-Credit with high market share in 2015 (Baseline Revenue) grew the fastest (YoY Pct Change) in in-app monetization, contributing to growth in the platform revenue from 2015 to 2016.*

Interviewer Comments

The candidate proposed grouping country and payments by baseline revenue and YoY Pct change to analyze revenue drivers. The approach can certainly highlight key drivers that contributed to the revenue changes.

> **#2 -** Suppose that a linear model was fitted on categorical variables, resulting in 0.99 R-squared. Should you use this model?

[Candidate] I'd like to ask questions first for clarification Is the response variable revenue? If so, is the analysis similar to the previous question dealing with understanding revenue drivers?

[Interviewer] Yes. This time a statistical model is applied to evaluate drivers and, I'd like you to focus on addressing the statistical condition described in the question.

[Candidate] To address this question, I need to first interpret the 0.99 R-squared value of the linear model. I'm going to assume that an ordinary least squares model was fitted on one-hot encoding of the categorical variables with high cardinality (i.e. country). When the number of predictors increases, R-squared becomes inflated because the model overfits under high-dimensionality.

[Interviewer] That's a valid assumption. How would you address this?

[Candidate] First of all, I would recommend not using R-squared soley. I would also use R-squared adjusted as the metric becomes penalized as predictors increase. Secondly, I would propose one of two changes: (1) Apply data grouping (2) Mixed modeling

[Interviewer] Can you elaborate on the approaches?

[Candidate] As I pointed out, the underlying problem is high-dimensionality of the feature set. Grouping each categories based on similar information can reduce the dimensionality. For instance, the country variable with 180+ values could be grouped into three continental regions - AMER, EMEA and APAC. This will reduce the R-squared inflation.

Another approach is to just use a mixed GLM which allows random slopes and intercepts of the categorical variables without increasing degrees of freedom.

Interviewer Comments

The candidate demonstrated a strong sense in statistical modeling methodology. His assessment of 0.99 R-squared was valid. The increased dimensionality of the feature set can cause R-squared to become high. The correction is either reduce the dimensionality with feature engineering or use a different model. In his response, the candidate outlined both, demonstrating solid knowledge in statistical modeling.

Interviewer Assessment

In the statistics section, a candidate is assessed based on correctness and soundness of statistical methodology, business sense and communication. For each dimension the candidate is rated in the following scale: (5) superior, (4) good, (3) adequate, (2) marginal, (1) not competent.

Assessments	Rating	Comments
Statistical Methodology	5	The candidate demonstrated strong competence in statistical methodology. In both problems, his responses were sound. In the first problem, he proposed the idea of segmentation using custom indexes to address business questions on key drivers in the Facebook games platform. In the second problem, he interpreted the 0.99 R-squared value correctly and proposed valid approaches. The sound approaches demonstrate that the candidate possesses a strong aptitude in statistical methodology.
Business Sense	5	The first problem assessed the candidate's business sense. The candidate demonstrated strong business sense as he understood the open-ended business problem and proposed a methodology that can work. He proposed bucketing drivers based on custom indexes to address business question such as which drivers contributed revenue growth on the Facebook games platform.
Communication	5	The candidate's responses were overall solid. He illustrated his points with tables and explained in an easy-to-follow manner.