GCI最終課題　README

締　切	2023/1/31 (火) 23:59
提出先	omnicampusから提出してください

1.課題 **2**
1-1.状況設定と出題 2
1-2.要件 2
1-3.出典の明記について 3

2.提出方法 **3**

3.ヒント **4**
目的変数について 4
事業提案について 4
市場分析について 4
探索的データ分析（EDA）について 4
モデル構築結果の提示について 4
説明性について 4
可視化について 4
何を求められるのか？ 5
事業提案の流れについて 5
なかなかイメージがつかめない方へ 5

4.データの内容 **5**
4-1.データベースについて 5
4-2.データベースの利用 5
4-3.カラムの説明 6

1.課題

1-1.状況設定と出題

あなたは、ITコンサルティング会社の若手アソシエイトです。
先日、チームのマネージャーが、とある電気通信事業会社A社からデータ分析のPoC(★)を受注してきました。
聞くところによると、A社には機械学習に詳しい人材はおらず、長年に渡り蓄積したデータをどのように事業に活かせばよいのかを苦慮しているとのことです。
今回のPoCでは、提供されたデータを機械学習を用いて分析することで、A社の課題を設定し、これを解決する新規事業の内容・効果を示します。
アソシエイトであるあなたは、約2ヶ月後のA社とのミーティングで、データ分析と機械学習に基づく新規事業を提案するプレゼンをしなければなりません。
分析と提案次第では、A社と正式にコンサルティング契約を結ぶことができ、あなたは、機械学習を用いた事業開発を主導することができるでしょう。

★ PoCとは、新しい技術・アイデアを完全に実用化する前に、それが実現可能であるか、目的が達成できるかなどの要件を確認するための実験的な検証のことです。PoCの結果、提案が要件を満たす場合に、実用可能なシステムの開発や詳細な事業戦略の策定にステップアップできます。

【出題】状況設定を踏まえて、事業モデルについてのプレゼンテーション資料（**15**スライド以内）を提出してください。

1-2.要件

資料作成にあたって以下の要件を定めています。作成にあたっては各回授業で修得した内容を活かすことを念頭におき、各要件を満たしてください。これらは提出物の評価項目でもあります。

要件を一定以上満たしていないものは不合格判定となる場合があります。提出前にも再度確認し、要件を満たした提出物となるよう留意してください。

1. 技術説明だけに終始せず、事業提案を行っていること
2. 一般論のみに基づく事業提案とせず、与えられたデータセットに対する探索的データ分析（**EDA**）を行うことで**A**社固有の課題を提示していること
3. 機械学習・統計分析モデル構築の結果を示し、スコアも明記していること（※下記「注意点」参照）
4. 外部データや論文等を参照して市場分析を行っていること
5. 機械学習に関して事前知識をもたないA社役員でも理解できるよう説明性を持たせてあること
6. 数値を羅列するだけに留まらず、データの可視化を行っていること

（※注意点）

現実の「機械学習について知らないA社役員向け」のプレゼンではあまり細かいモデルの話はしても仕方ないという判断もありえますが、この最終課題はあくまでGCIで講義した各回の内容を受講生が活用できるようになることを目的に出題するものですので、この趣旨に基づき要件3の「機械学習・統計分析モデル構築の結果」についてはどのような取り組みを行ったか意識して具体的に記述してください。（なぜそのモデルを選んだか、どのような構築内容としたか、複数モデルの比較検証の過程、スコアなど）

1-3.出典の明記について

分析や提案を思いつきに終わらせないことを目的に一般公開されているデータや手法を引用すること自体は学術研究における先行研究調査と同様奨励されるべきことですが、その一方で出典を明記せず他者の取組を引用する剽窃行為が残念ながら後をたちません。

技術者倫理を全うすることは知識や技術以前の絶対条件であり、GCIでは不正行為に対して厳しい対応をとっています。引用が剽窃行為とならないよう、以下の点にはぜひ留意してください。

1. 外部のデータ・分析手法・図表・文章等を引用する場合には、必ず出典を明記してください（出典なき引用は剽窃と判断します）。
2. コピー＆ペーストそのものでなくても、参照元と同じ手法で分析や可視化を行う場合には出典を明記する必要があります。（過去に少なからぬ受講生がほぼ同じ可視化をしていたことがありましたが、このような可視化を成果として評価対象とすることはできません。）
3. データ分析とモデル開発においては、誰でも無償でアクセス可能なオープンデータを追加で用いても構いません。その場合にも、出典を明記してください。
4. データ分析とモデル開発以外に、独自の市場調査をプレゼン資料に含めても構いません。データを引用する場合には、上記と同様に出典を明記してください。

※引用の分量について

引用はあくまでそれを土台に自分自身の新しい知見を展開するために行うものであることに留意してください。引用は論述の展開に資する一方、引用箇所自体が評価対象となるわけではありません。過去にはプレゼンテーション全体がほぼ引用のパッチワークに終わってしまっている事例も見受けられましたが、このような場合には提出者自身の取組があったとは評価しかねることになります。

目安として、引用そのものはプレゼン全体の概ね40%以下に留めることが望まれます。「Telecom Custamer Churn」のデータセットは一般公開されていますが、特にこれの解析を目的とする資料（可視化・ロジック等）や過去の最終課題をそのまま引き写している部分は一切評価対象になりません。ただし、それらの資料を基に自身の考察やオリジナリティのある解析を追加で行う場合は、その差分を明確にしたうえで用いることは構いません。

2.提出方法

提出先	omnicampus
締切	2023/1/31 (火)　23:59
形式	PDF形式 ただし、PowerPointやKeynote等のプレゼンテーションツールを用いたスライドで作成してください。
分量	**15**スライド以内
ファイル名	"[omnicampusアカウント名].pdf"としてください。(例) アカウント名がABCの場合→ "ABC.pdf"
ファイルサイズ	10MB以下
再提出する場合	提出期間中に複数回提出し直すことが可能です。複数回提出した場合には、最新の提出物のみが採点対象となります。
修了判定	修了判定結果を通知する日程は、締切後にSlackで連絡します。
優秀修了生	最終課題の達成内容にコンペの成績を加味し、特に優れた受講生を「優秀修了生」として選出します。 優秀修了生の発表日程も締切後に別途Slackで連絡します。

3.ヒント

目的変数について

この最終課題では目的変数は自由に設定していただくことができます。

EDAや市場分析を通じて見えてきた課題に応じて、何を目的変数とするか考察してください。（奇をてらうこと自体には意味がありませんので、クライアント企業の課題が何であるかという問題設定に基づいて定めることを重視してください。）

事業提案について

A社は、あなたが通常のビジネスコンサルではなくデータサイエンティストであるからこそ今回のPoC案件を委託しています。このことを意識し、一般的なビジネス論だけでなく、機械学習・統計分析モデルに基づき事業提案することが必要です。

事業提案が実現性のないアイデアに終わらないために、データサイエンスの知識を駆使して事業規模・効果の定量評価を行うことも説得性を高めます。難しいことですが、斬新性や新規性があればなお高い評価に繋がるでしょう。

またSQLを活用した事業提案、特にリレーショナルデータベースでしかできないことに着目した提案・考察も評価対象となります。

市場分析について

ドメイン知識はデータ分析にとって重要な位置を占めています。外部のデータや論文・記事等をサーベイして市場分析を行い、さらに仮説を示すことができれば、事業提案の妥当性に対する説得力を高め、評価につなげることができます。

探索的データ分析（EDA）について

A社にとって貴重な顧客データを委ねられていることを重く受け止め、外部データの引用だけに終わらせず与えられたデータセットそのものに対してEDAを行い「このデータセットから見えてくること」を提示することが重要です。そのためには問題提起に資する特徴的なデータ分布を発見して提示し、そこから見えてくるA社固有の課題を論理の飛躍なく示すことが必要です。

さらに、A社はデータサイエンティストを抱えておらず、データセットの構成が必ずしも合理的であるとは限りません。特徴量エンジニアリングを行う技術もあなたに期待されていることであり、有効な内容を提示すれば評価対象となります。特徴量エンジニアリングが予測精度向上を帰結すればモデル評価にとってもプラスになり得ます。

モデル構築結果の提示について

本件はPoCであり、ある程度詳細な技術情報の提示を行う必要があることに留意してください。A社が持つ顧客データの傾向性やあなたの問題意識に適合するように複数のモデルを作成し、各モデルの仕組みや理論的背景・長所/短所を踏まえて比較を行い、明確な理由をもってモデル構築（アンサンブル等も含む）を行い、モデル選定の理由や出力結果に対する考察を説明できていることが評価に繋がります。

A社がより性能の高いモデルを有する競合他社に流れてしまわないよう、スコアをアピールすることも大切です。その際、課題設定に合った「適切な評価指標」を活用できているか否かも評価対象となります。また、ベースラインモデルも構築してこれとの比較によってスコアを示せば、より具体的にあなたの作成したモデル性能の高さを訴えることができます。

PoCはコンペと異なりモデルを作って終わりではないため、モデルを通じて事業提案を補強しうる仮説を示すことも重要な点です。

説明性について

作成した資料は機械学習に関して事前知識をもたないA社役員に対するものです。そのため専門用語の説明なき濫用は説明性の欠如とみなされます。その一方で、本件はPoCでもあるため、技術・実装の説明を省くのも望ましくありません。重要なことは実装した内容や用いる技術について非エンジニアでも妥当性が納得できるような言葉に置き換え、適宜概念図なども示しながらわかりやすく説明することです。

メッセージや各スライドタイトルを明確にしたり、適宜サマリを配置するなどちょっとした工夫が相手に伝わるプレゼンテーションを可能にします。

可視化について

なんとなくの可視化で終わらせないために、配色や凡例表示、単位表記やフォントなど、図表を構成する各要素について「なぜこのようなグラフにしたのか」明確に理由を自分で説明できるような仕方で図をつくることが相手にとって見やすい可視化につながり、高い評価を勝ち得る要素となります。文字がつぶれていたり、外れ値の処理やビニングを怠って極端に偏ったグラフなどは誰にとっても見づらく、良い印象を残しません。主張を明確に伝えることができるよう、目的意識をもって図表設計を行いましょう。

何を求められるのか？

与えられたデータから企業が持つ課題を明確化し、収益を上げるための事業を考えます。その際に評価されるのは、的を射た問題設定と事業提案の必然性の論理、そして事業規模の適切な推定です。

事業提案の流れについて

事業提案の流れについてイメージがわかない場合には、まずは以下のような形で「1-2.要件」の各項目を確実に満たしていくように流れを素描してみることを推奨します。

- 導入
    1. 市場分析
        i. 外部データの参照
        ii. 考察
- SQLによるデータ取得
- EDA
    1. 可視化
    2. 考察
- 問題設定
- 特徴量エンジニアリングの実施
- 実験結果
    1. モデル選定・評価
    2. 可視化
    3. 考察
- 事業提案
    1. 事業案の提示
    2. 導入規模・効果の定量評価

なかなかイメージがつかめない方へ

イメージがつかめない方は下記の資料も参考にしてください。

https://docs.google.com/document/d/1BouUUlToQi_7Qbm301-BmPmnu9jvoGDFizC3Sj-JmuY/edit?usp=sharing

4.データの内容

4-1.データベースについて

- 電気通信事業会社A社の過去のデータが含まれたデータベースファイル「telecom.sqlite3」が配布されている
- カラム名とその説明が明らかになっている（「4-3.カラムの説明」参照）
- 以下の情報が格納された2種類のテーブルが含まれている
    1. 顧客情報（Client）
    2. 利用履歴（Record）

4-2.データベースの利用

A. SQLiteのインストールができている場合
    1. SQLiteを用いて、データベースに接続
    2. 各自、データベースを観察
    3. データ分析、事業提案に必要なデータをcsv形式で出力
    4. 各自必要なpythonモジュール（pandasなど）でデータ分析

B. SQLiteのインストールができていない場合
    1. pythonの標準ライブラリを用いて、Aと同様のことを実行する

C. 何から手をつければ良いか分からない場合
    1. 参考資料として配布されているサンプルコードを使用する
    2. 全データをcsv形式で出力
    3. 各自必要なpythonモジュール（pandasなど）でデータ分析

4-3.カラムの説明

A社の担当者にデータについて問い合わせたところ、以下の説明が送られてきました。現段階では正式なコンサルティング契約は未締結であり、情報保護規程の厳しいA社からこれ以上の詳しい情報開示を受けることはできません。

カテゴリカル変数の各種記号や略語の意味など不明な点は多々ありますが、不明な点があること自体も業務上の与条件として分析を進めてください。

カラム名	内容の説明
rev_Mean	Mean monthly revenue (charge amount)
mou_Mean	Mean number of monthly minutes of use
totmrc_Mean	Mean total monthly recurring charge
da_Mean	Mean number of directory assisted calls
ovrmou_Mean	Mean overage minutes of use
ovrrev_Mean	Mean overage revenue
vceovr_Mean	Mean revenue of voice overage
datovr_Mean	Mean revenue of data overage
roam_Mean	Mean number of roaming calls
change_mou	Percentage change in monthly minutes of use vs previous three month average
change_rev	Percentage change in monthly revenue vs previous three month average
drop_vce_Mean	Mean number of dropped (failed) voice calls
drop_dat_Mean	Mean number of dropped (failed) data calls
blck_vce_Mean	Mean number of blocked (failed) voice calls
blck_dat_Mean	Mean number of blocked (failed) data calls
unan_vce_Mean	Mean number of unanswered voice calls
unan_dat_Mean	Mean number of unanswered data calls
plcd_vce_Mean	Mean number of attempted voice calls placed
plcd_dat_Mean	Mean number of attempted data calls placed
recv_vce_Mean	Mean number of received voice calls
recv_sms_Mean	NaN
comp_vce_Mean	Mean number of completed voice calls
comp_dat_Mean	Mean number of completed data calls
custcare_Mean	Mean number of customer care calls
ccrndmou_Mean	Mean rounded minutes of use of customer care calls
cc_mou_Mean	Mean unrounded minutes of use of customer care (see CUSTCARE_MEAN) calls
inonemin_Mean	Mean number of inbound calls less than one minute
threeway_Mean	Mean number of three way calls
mou_cvce_Mean	Mean unrounded minutes of use of completed voice calls
mou_cdat_Mean	Mean unrounded minutes of use of completed data calls
mou_rvce_Mean	Mean unrounded minutes of use of received voice calls
owylis_vce_Mean	Mean number of outbound wireless to wireless voice calls
mouowylisv_Mean	Mean unrounded minutes of use of outbound wireless to wireless voice calls
iwylis_vce_Mean	NaN
mouiwylisv_Mean	Mean unrounded minutes of use of inbound wireless to wireless voice calls
peak_vce_Mean	Mean number of inbound and outbound peak voice calls
peak_dat_Mean	Mean number of peak data calls
mou_peav_Mean	Mean unrounded minutes of use of peak voice calls
mou_pead_Mean	Mean unrounded minutes of use of peak data calls
opk_vce_Mean	Mean number of off-peak voice calls
opk_dat_Mean	Mean number of off-peak data calls
mou_opkv_Mean	Mean unrounded minutes of use of off-peak voice calls
mou_opkd_Mean	Mean unrounded minutes of use of off-peak data calls
drop_blk_Mean	Mean number of dropped or blocked calls
attempt_Mean	Mean number of attempted calls
complete_Mean	Mean number of completed calls
callfwdv_Mean	Mean number of call forwarding calls
callwait_Mean	Mean number of call waiting calls

churn	Instance of churn between 31-60 days after observation date
months	Total number of months in service
uniqsubs	Number of unique subscribers in the household
actvsubs	Number of active subscribers in household
new_cell	New cell phone user
crclscod	Credit class code
asl_flag	Account spending limit
totcalls	Total number of calls over the life of the customer
totmou	Total minutes of use over the life of the customer
totrev	Total revenue
adjrev	Billing adjusted total revenue over the life of the customer
adjmou	Billing adjusted total minutes of use over the life of the customer
adjqty	Billing adjusted total number of calls over the life of the customer
avgrev	Average monthly revenue over the life of the customer
avgmou	Average monthly minutes of use over the life of the customer
avgqty	Average monthly number of calls over the life of the customer
avg3mou	Average monthly minutes of use over the previous three months
avg3qty	Average monthly number of calls over the previous three months
avg3rev	Average monthly revenue over the previous three months
avg6mou	Average monthly minutes of use over the previous six months
avg6qty	Average monthly number of calls over the previous six months
avg6rev	Average monthly revenue over the previous six months
prizm_social_one	Social group letter only
area	Geogrpahic area
dualband	Dualband
refurb_new	Handset: refurbished or new
hnd_price	Current handset price
phones	Number of handsets issued
models	Number of models issued
hnd_webcap	Handset web capability
truck	Truck indicator
rv	RV indicator
ownrent	Home owner/renter status
lor	Length of residence
dwlltype	Dwelling Unit type
marital	Marital Status
adults	Number of adults in household
infobase	InfoBase match
income	Estimated income
numbcars	Known number of vehicles
HHstatin	Premier household status indicator
dwllsize	Dwelling size
forgntvl	Foreign travel dummy variable
ethnic	Ethnicity roll-up code
kid0_2	Child 0 - 2 years of age in household
kid3_5	Child 3 - 5 years of age in household
kid6_10	Child 6 - 10 years of age in household
kid11_15	Child 11 - 15 years of age in household
kid16_17	Child 16 - 17 years of age in household
creditcd	Credit card indicator
eqpdays	Number of days (age) of current equipment
Customer_ID	NaN