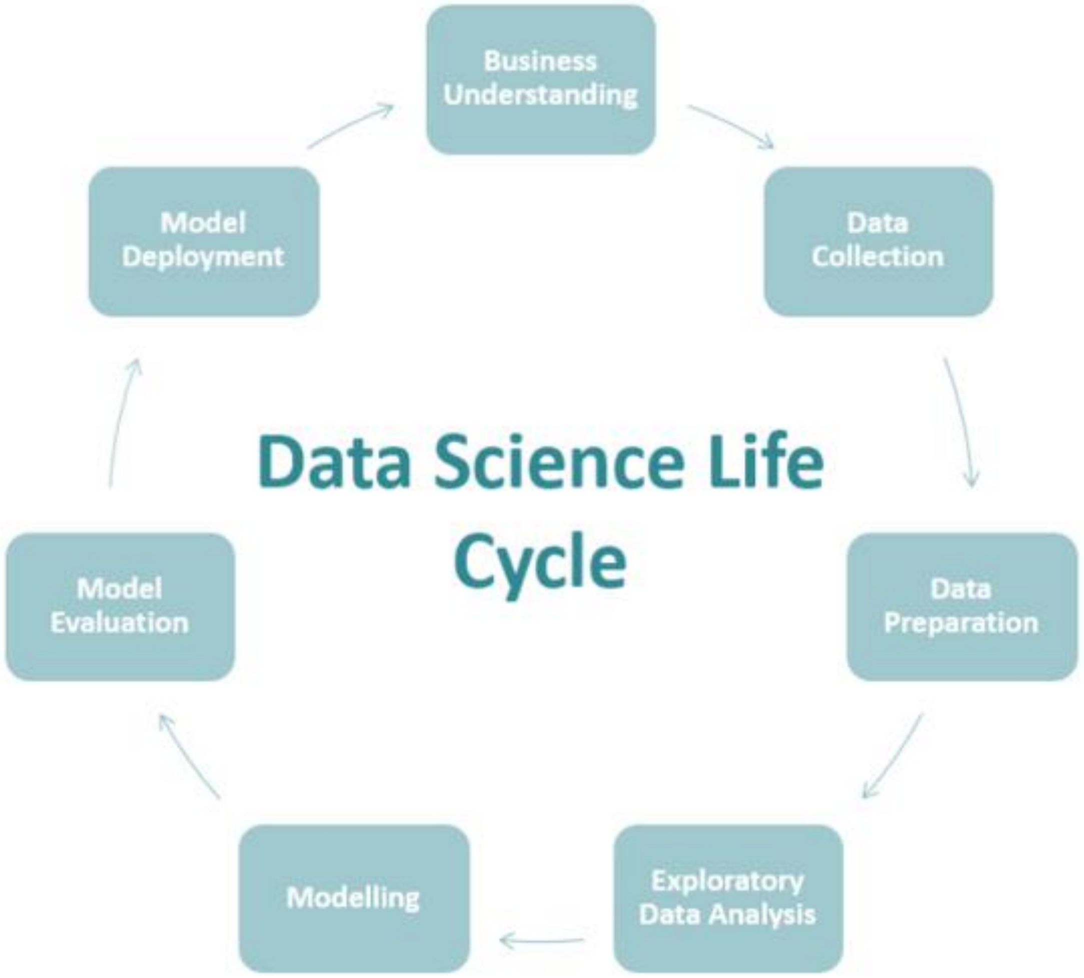
Analysis on ABC LLC’s customers

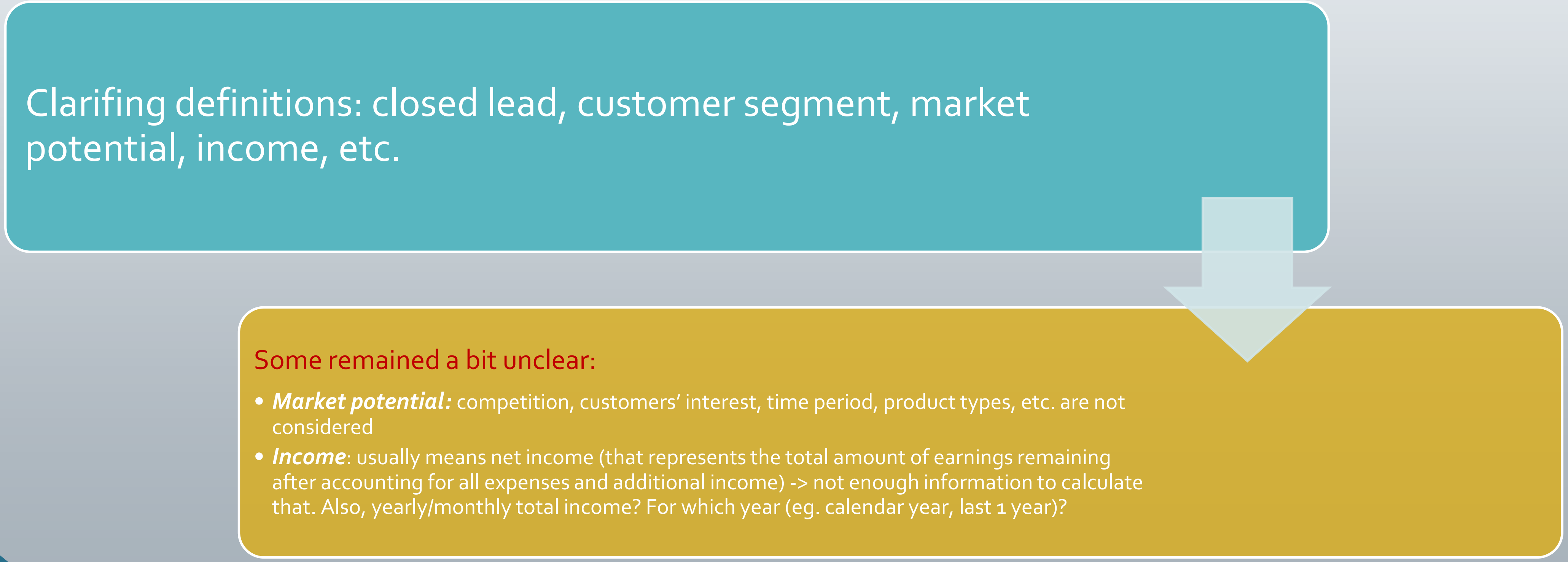
Task from boberdoo.com

Questions

- Considering the information on closed leads, how would you describe Abc LLC's current most typical client?
- Considering the information on closed leads, what customer segments should Abc LLC target to maximize their income?
- How would you describe the current market potential for the segments from Question 2?
- How much more should Abc LLC bid on average on the segments from Question 2 to increase their income by 30%? Would the investment pay off?


Data Science Life Cycle
Business Understanding
Data Collection
Data Preparation
Exploratory Data Analysis
Modelling
Model Evaluation
Model Deployment

Business understanding



Clarifing definitions: closed lead, customer segment, market potential, income, etc.

Some remained a bit unclear:

- ***Market potential:*** competition, customers' interest, time period, product types, etc. are not considered
- ***Income***: usually means net income (that represents the total amount of earnings remaining after accounting for all expenses and additional income) -> not enough information to calculate that. Also, yearly/monthly total income? For which year (eg. calendar year, last 1 year)?

Data collection, preparation and exploratory analysis I.

Steps followed:

1. Loading data and merging into one dataframe
2. Data exploration/univariate analysis: checking for NaNs, outliers and distribution
3. Creating categorical variables (age, income) for analysis on typical client
4. Converting date variable to Timestamp for better handling
5. One-hot-encoding[1]: introduction of binnary variables created from categorical features
6. Standardization: MinMaxScale[2] and RobustScaler[3]: considering outliers and binnary variables

[1] Encoding categorical variable to binnary (0-1 values) variables. The variables number is equal to or one less than the levels of the categorical variables

[2] For each value in a feature, MinMaxScaler subtracts the minimum value in the feature and then divides by the range. It doesn't reduce the importance of outliers.

[3] RobustScaler transforms the feature vector by subtracting the median and then dividing by the interquartile range (75% value — 25% value). It reduces the effects of outliers, but it does not scale the data into a predetermined interval like MinMaxScaler

Data collection, preparation and exploratory analysis II.

Problems

Additional information about the possible leads

- What other factors have an effect on
  - successfull bidding,
  - selling health insurance and
  - contract size?
- Info helping the evaluation of customers' health condition and financial background could have been valuable (may raise ethical issues)

Handling possible outliers:

- With respect to household income and contract size, more information would have been valuable to decide upon elimination of outliers

Binnary variables:

- Introducing binnary variables made clustering more challenging

Analysis/Modeling I. Most typical customer

- Finding the 5 most typical customer groups based on filtering on demographics, closed leads only:
  - **Filters**: state, gender, age_c, income_c; Calculating average contract size for every group
  - ***The typical client of ABC LLC is a young adult male who lives in New York, whose household income belongs to the lowest segment of the sample and average contract size is around 740 dollars.***

Analysis/Modeling II. Customer Segmentation I.

Hierarchical clusters:

- **Cluster 1:** Mixed men and women (with more men), older adults, with middle HH income, having average contract size the highest, mostly from NY, MA, WA
- **Cluster 2:** Exclusively men, middle aged, their HH income is around the median, having contract size above the median, mostly from MA, WA, NJ
- **Cluster 3:** Exclusively women, middle aged, from low income HH, having contract size around the median, mostly from NY, WA, MA
- **Cluster 4:** Exclusively men, youngest average age, more HH in the low income group, having contract less than average, mostly from NY, IL, OH

Targeting strategy/Opportunities:

- First and fourth clusters: the lower number of successful sales there
- First cluster's high average contract size
- Other states like NJ and CT are present in case of more than one segment.
- Gender specificity can be recognized in the sample, meaning that gender based targeting can be useful.
- Number of successful sales in the younger and older adult groups may be even higher

Analysis/Modeling II. Customer Segmentation II.

Problems

K-means algorithm provides unsatisfactory results:

- Cluster characteristics highly depend on encoding and standardization method
- K-means is for minimizing the within-cluster squared Euclidean distances between the clustered observations and the cluster centroid.
- Therefore, it should only be used with data where squared Euclidean distances would be meaningful.

Hierarchical clustering is a bit arbitrary:

- It requires a decision where to 'cut the tree' to get the final cluster assignments
- Some domain and business knowledge is essential to determine the right number of clusters

Other Algorithms:

- Considering mixed data type, other algorithms should be tried out as well, like Density-based spatial clustering of applications with noise (DBSCAN) or neural networks

-> Amir Ahmad and Shehroz Khan: Survey of State-of-the-Art Mixed Data Clustering Algorithms. IEEE Access (Volume: 7), link: https://arxiv.org/pdf/1811.04364.pdf

Analysis/Modeling III. Market Potential

Segment 1:

- Customer base: 5371
- Estimated potential income based on average contract size : 9179039.0 USD

Segment 2:

- Customer base: 19379
- Estimated potential income based on average contract size : 18003091.0 USD

Segment 3:

- Customer base: 30181
- Estimated potential income based on average contract size : 28822855.0 USD

Segment 4:

- Customer base: 19959
- Estimated potential income based on average contract size : 16286544.0 USD

Problems with estimating market potential

- No information on competition, customers' interest, time period, product types, etc.
- Overlaping segments make hard to calculate total market potential

Analysis/Modeling IV.
Estimating investment for increasing yearly income I.

Univariate regression:

- according to the p value, the relationship between the amount of bid and contract size is significant.

Coefficient:

- The coefficient of 3.0561 means that as the max bid variable's unit increases by 1, the predicted value of contract size increases by 3.0561.

Increasing spend on bids:

- ABC LLC should increase spend on bidding around 42396 USD in 2019 to meet 30% increase in income. The investment would pay off.

Analysis/Modeling IV. Estimating investment for increasing yearly income II.

Problems

Winning a bid was not addressed in the univariate model, while contract size depends highly from this:

- Win: contract size >= 0
- Lose: contract size = 0
- But: Multicollinarity with bids are present
- Possible solution: **Two-Stage Regression Analysis or Partial least squares regression.**
- **Casual relationships: bids -> win, win -> contract size**

Estimating the effect of other variables (holding them constant) on income is a must:

- Product(s) characteristics
- Customer characteristics
- Competition