Case Study Introduction – Data Science

The accompanying attachments contain two years of season ticket renewal observations for the local basketball team. Since the team's ticketing department is separate from the customer service department, they collect data in different spreadsheets and they need your help to combine renewal information with customer data to gain a better understanding of why some season ticket members aren't renewing. Luckily, both departments record a CustomerID and a Renewal (or NonRenewal) Date as unique identifiers across both data sets. But be warned…the departments don't always record this data in the same format! Below is a quick overview of the data:

Renewal Details

- Column A: Customer ID.
- Column B: The previous season for this observation.
- Column C: The renewal season for this observation
- Column D: Indicator of whether the customer renewed or not for the renewal season (1= renewed).
- Column E: The date on which a customer renewed. If they did not renew, then this date represents the renewal deadline (which passed without the Customer renewing her/his seats).
- Column F: Average Price Paid Per Seat for Season Tickets in the Previous Season
- Column G: Total season tickets purchased in previous season
- Column H: Team's winning % in previous season
- Column I-K: Home games attended, missed (not attended), and total home games from the previous season
- Column L: Non-game events, such as "meet the players" or other special events for season ticket holders, attended by the customer in the previous season.
- Column M: Phone calls made to season ticket holder in the last 90 days before their renewal date (simple count regardless of whether customer picked up). For those that did not renew, this pertains to the non-renewal deadline date.
- Column N: Live phone conversations had with customer in the last 90 days before their renewal /non-renewal date.
- Column O: Emails sent to customer in the last 90 days before renewal/non-renewal date.
- Column P: Generalized section in which customer sat in the previous season.

Customer Details

- Column A: Customer ID.
- Column B: The date on which a customer renewed. If they did not renew, then this date represents the renewal deadline (which passed without the Customer renewing her/his seats).
- Column C: Customer's distance from stadium based on zip code.
- Column D: Tenure as season ticket member through previous season (e.g., Row 2 was a first-year season ticket holder in 2012, hence they have a tenure of 1).

Case Study Questions

Using this dataset, please build a retention model for the local basketball team. To help guide your thinking, we've listed the six steps in the Data Science Project Lifecycle. Please answer the following questions, designed to understand your general approach to data and modeling.

1. **Discovery**

   Describe the general approach and methodology you would use to build a likelihood to renew model from this data set. What would the ultimate objectives of the model be? What type of algorithms/models or analytical tools would you consider using?

2. **Data Preparation and Initial Data Analysis (IDA)**

   Since the majority of this data has been gathered for you, the next step is to prep and clean the information by handling NAs/nulls/missing values and identifying/correcting potential errors. Please complete this step in Python, ***please provide code with comments****. Part two of the Data Science Project Lifecycle also includes identifying any additional data needed. For example, joining the Customer Details file to the Renewal Details file would add a few additional data points about the customer to the dataset. Additionally, please list a few examples of data points* ***not included in the provided data*** *that might help predict customer retention*

   - *While this data set is mostly "clean", what are some of the challenges you could see in creating this data set from scratch? No need to list more than a few of examples.*

3. **Model Planning and Exploratory Data Analysis (EDA)**

   This is the time to get familiar with your data by poking around and building visualizations that help you understand the relationships between variables. Please provide the most important visualizations and/or summary statistics that you used to help guide your approach to variables selection and feature engineering decisions.

4. **Model Building**

   Using Python, build a retention model and ***provide your code*** *and* ***any files you used****. It is best to test and validate multiple models for comparison. There is no one right answer, so feel free to be creative and try new techniques.*

5. **Communicate Results**

   A famous statistician once said, "All models are wrong, but some are useful." While this is true, you can take it one step further to say that even the best model in the world will not be useful unless you can accurately communicate your findings and gain buy-in from key stakeholders. ***Please develop a PowerPoint deck with the appropriate charts and graphs*** *to explain your model, key metrics and findings, etc. Please include actionable business recommendations on how the executive team can use the findings from the model to improve customer retention next year.*

6. **Operationalize**

   Now that we've successfully delivered the model to the leaders for our local basketball team, they want the model to be run in production on a recurring basis. ***Describe how you would operationalize and deploy the model. Include details about the tools, frameworks, and best practices you would use for model deployment in a production environment.*** *How would you monitor and maintain the model over time?*