**Brisa H, Njenga G, Rashmi K**

**10/11/2025**

**DATA 5100**

**Customer Lifetime Value: Modeling Customer Profile for Underwriting Optimization**

**Problem statement:** Predict the customer lifetime value based on the customer profile, to decide whether a prospective customer would be a profitable policy to write

**Data sources:** We will use the [IBM Watson Marketing Customer Value Data](https://www.kaggle.com/datasets/pankajjsh06/ibm-watson-marketing-customer-value-data/data) to create our predictive model. We will not need to impute any of the data since all of it is present, however we will want to perform basic feature engineering. We have a few known limitations to our data; first it is limited to the West Coast only. The states included are Arizona, Washington, California, Nevada and Oregon. We have approximately 9,000 customer profiles available to us to model with, and we’ll need to determine if the data is skewed or if our sample is an appropriate measure of the population. Lastly, the data is based on policies effective through 2011. We will also continue to look for more updated or relevant sample data to use instead of, or as well as.

**Analytical approach:** We will start by taking a descriptive approach during the exploratory data analysis to better understand the relationship between all of our data and variables. Any diagnostic analysis would be focused on the correlations in our dataset, and testing any relationships present to confirm predictive power. We plan to take a predictive approach as we want to use our model to predict customer lifetime values for new customer profiles. Ultimately we plan to make recommendations as to whether or not to keep a customer, therefore we will end with a prescriptive approach.

**Solution technologies:** We plan to use Python 3, Github, Jupyter Notebooks and do not believe we will require any additional computational power to answer our business problem.

**Challenge:** We know that the data is older and potentially not as relevant as data from more recent policyholders, furthermore we are not sure if we can extrapolate findings from the West Coast to the United States as a whole. We are also concerned that our sample size may not be representative of the whole population, even for some of the states where we have data due to large populations within them (California).

**Citations:** To be determined as we work the problem.

**Group Dynamics:** Weekly progress updates, to-do lists and task delegation will be documented on both our Group 4 canvas page and the Git Hub. All other communications will be done using SeattleU email IDs. To maintain clarity, we will continue conversations within the same email thread whenever possible. We will share data and code using Git Hub. Because we are spread out geographically, and have various commitments outside of school we have agreed to set meeting times and deliverables as needed, rather than on a recurring on-going basis. We will often meet virtually via zoom, but prioritize in person meetings for high importance action items.

**Next Steps:**

* Brisa to set up our [Git Hub](https://github.com/brisamh/modeling_clv) and enable permissions for Group 4
* All to download data and begin preliminary EDA, next **meeting on 10/17 11am at SeattleU campus** to review our findings and proceed with a plan for the modeling or any additional deliverables that arise from our findings.