Nutritional Labels for Automated Decision Systems

DS-UA 202 Responsible Data Science Project

May 09, 2022

Alexis Martin

Michael Shu

I. Background

Data science is highly prevalent in making predictions in the insurance industry. The Allstate Purchase Prediction Challenge(APPC) seeks to shorten the insurance quoting process by predicting a customer's purchase sooner in the shopping window, in turn, lessening the likelihood that the issuer will lose the customer's business. In short, Allstate is looking to acquire an ADS that predicts a customer's coverage options. Understanding the importance of being mindful of protected attributes when curating such systems, our project pinpoints any avenues for bias and disparities amongst customers the ADS may unknowingly propagate.

The APPC ADS's primary goal is to predict whether an individual purchases an insurance policy or not, based on several customer attributes. Several of these attributes are protected and allow us to cast an interesting lens on whether substantial bias arises, affecting customers and their purchasing decisions. Such attributes include a customer's purchasing history, including previous insurance products and their respective quotes, selected coverage option purchased, the state that the insurance product and quote were viewed in, and whether the customer represents themself or a collection of people. In addition to these attributes homeowner and marital status, previous coverage and the duration, age and risk factor assessment score can be argued as protected attributes and will be treated as such in our analysis.

Although the ADS may be able to aid Allstate in offering competitive rates based on quotes from other policy providers / competitors, the system may discriminate against

age groups younger than 25 and older than 65, households without a married couple, and homeownership.

Since the risk score of an individual is often dependent on protected attributes that shouldn't be counted such as age and marital status we show that this ADS predicts individuals across protected groups differently, and therefore has disparate impact upon these groups. We show that recommending certain insurance plans leads to skewed underlying distribution in plans shown to customers.

Using a nutritional label system for determining the health and viability of the ADS from a responsible data science perspective, we will establish a focus on several quantitative bias signals to establish the benchline scores for how the APPC ADS performs. For this reason we will use correlation matrix, feature_importance, difference in distributions and XGBoost to analyze the influence of various protected attributes on accuracy scores.

II. Input and Output

Due to insufficient metadata about the data given, we can not say for certain how the data was collected. The only criteria that we can assume All-State selected data upon was whether or not a customer was viewing insurance options or not. The data provided by Allstate includes: customer ID, point of contact for shopping, indicator of shopping or purchase point, day, time, state, location, household size, homeowner indicator, car age, car value, risk factor score, age, whether the customer is married, current or former product options, duration of previous coverage and cost. Given insurance marketing

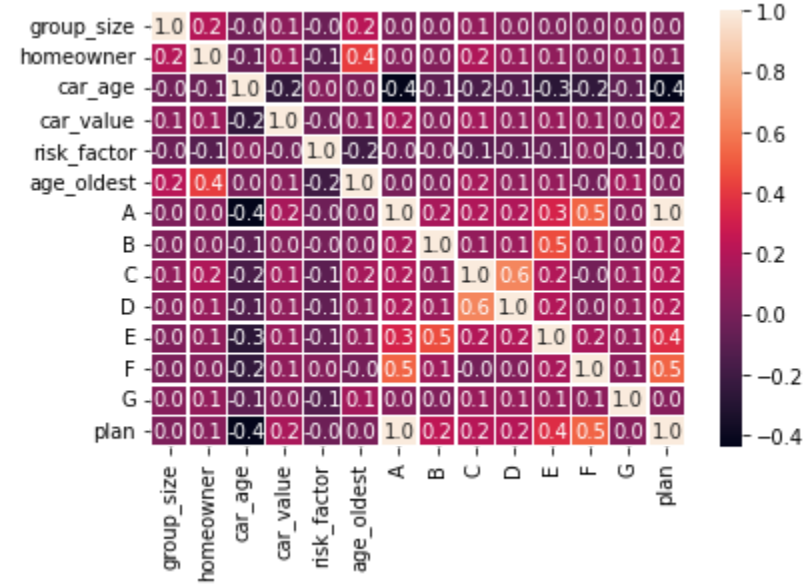
tactics we suspect that Allstate partners with comparative online insurance agencies that aggregate quotes for shoppers (i.e. thezebra.com). Using web browser cookies collected from several sites, Allstate can track potential customer's activity and interest in their products. Input features and their attributes are as follows:

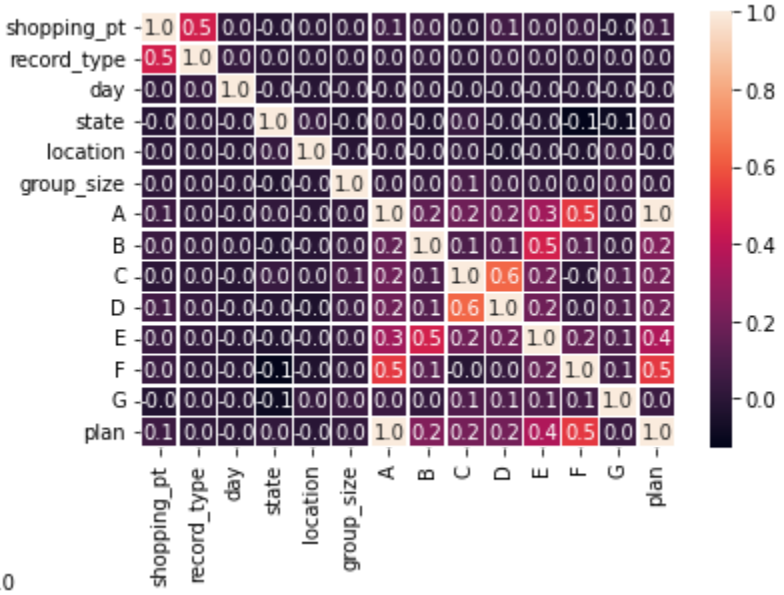
- **customer_ID:** is a unique identifier for the customers. It is an integer.
- **shopping_pt:** is a unique identifier for the shopping point of a given customer. Values range from 1-13. Indicates which partner Allstate aggregate insurance quoting website.* It is an integer.
- **record_type:** 0=shopping point, 1=purchase point. The value is binary. It indicates whether a customer is redirected to Allstate's checkout webpage.* It is an integer.
- **day:** Day of the week that the insurance policy was searched for. Range from 0-6, with 0 being Monday. It is an integer.
- **time:** Time of day that the insurance policy was searched for. 24 hour framework. A string taking on the form "HH:MM". eg: 08:35.
- **State:** State where shopping point occurred. A string consisting of two letters designating the state. eg: NY or CA
- **location:** Location ID where shopping point occurred. It is a float, ranging from 10001, to 16581.
- **group_size:** How many people will be covered under the policy (1, 2, 3 or 4). It is an integer.
- **homeowner:** Whether the customer owns a home or not (0=no, 1=yes). It is an integer.
- **car_age:** Age of the customer's car, presumably in years since it was not specified by AllState whether this was months or years. It is an integer
- **car_value:** How valuable was the customer's car when new. It is a string consisting of a single value from a,b,c,d,e,f,g,h or i. There is no specification on what these values mean.
- **risk_factor:** An "ordinal assessment" of how risky the customer is (1, 2, 3, 4). It is not specified as to how these values were calculated. It is a float.
- **age_oldest:** Age of the oldest person in the customer's group. We assumed "group" to mean all the members of the insurance plan. It is an integer.
- **age_youngest:** Age of the youngest person in the customer's group. It is an integer.
- **married_couple:** Does the customer group contain a married couple (0=no, 1=yes). It is an integer.
- **C_previous:** What the customer formerly had or currently has for product option C (0=nothing, 1, 2, 3,4). It is a float
- **duration_previous:** how long (in years) the customer was covered by their previous issuer. It is a float.
- **cost:** cost of the quoted coverage options. It is an integer.
- **A,B,C,D,E,F,G:** the coverage options - Since there was not sufficient metadata, we've decided to speculate on how to interpret this. Since the letters represent certain coverage options, we interpreted the numbers as representing how much coverage a certain plan has. For example, a value of 0 means no coverage, and 4 being the most coverage. Therefore, A,B,E and F are optional insurances, while C,D and G are not. For example, a driver must have a minimum insurance plan. A U.S. citizen must have a minimum health insurance plan, or they will be taxed. Attached below is a diagram provided by AllState, designating the insurance options and the possible values they can take on.

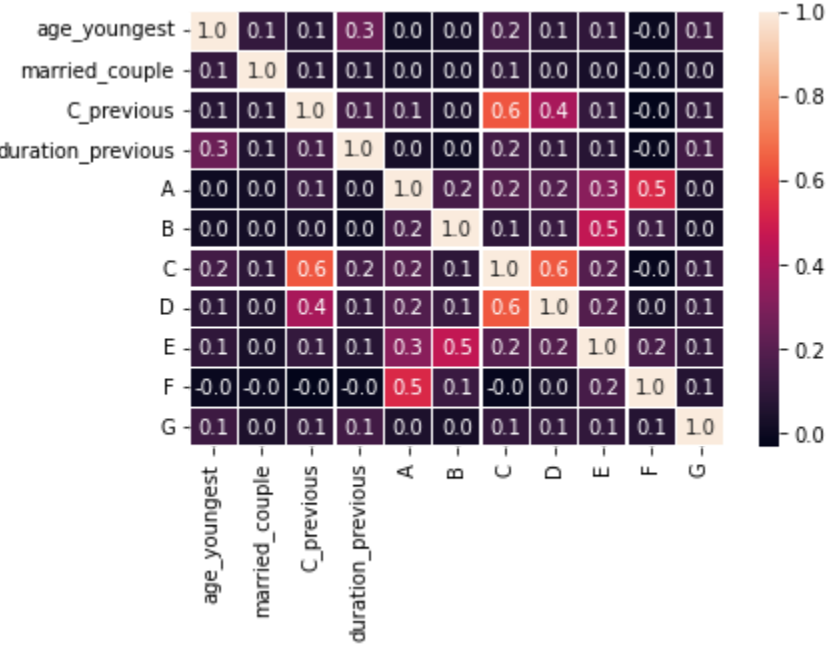
Option name	Possible values
A	0, 1, 2
B	0, 1
C	1, 2, 3, 4
D	1, 2, 3
E	0, 1
F	0, 1, 2, 3
G	1, 2, 3, 4

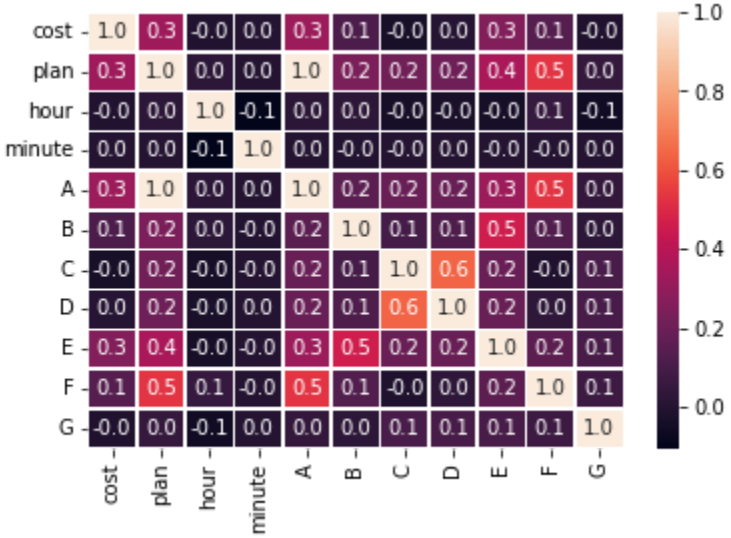
Attached below are several correlation graphs created within the ADS by the coder. Most are self explanatory and do not observe an extremely strong correlation between the values of the plans and any other variables. There is a notable correlation between C_previous and C, which makes sense since C_previous is data about what the person previously chose for their C plan. For example, if a person had homeowner's insurance, previously, it would be very likely that he/she would seek homeowner's insurance in their new quote. There is also the value of "plan" that we must explain. The ADS condensed the values of A,B,C,D,E,F, and G into a single value "plan" that is a 7 digit amortization of all the values of the plan options taken together. He did this by simply multiplying each value by a power of 10. So A times 10^6, B times 10^5, C times 10^4, etc. We can see in all the correlation plots that A is 100% correlated with plan, presumably since A received the largest change to its value at 10^6. There is also "hour" and "minute" which are both self-created by the ADS. They are the integer counterparts of the time string. For

example, if the time is “08:35”, the hour would be 8 and the minute would be 35.









III. Implementation and Validation

The ADS performs “fillna” on the train datasets variable “risk_factor”. The method of filling was set to “ffill”. This mode takes the value directly preceding a NAN to all the consecutive NANs following it. The ADS also performs this for the variables

“car_value”,”C_previous”, and “duration_previous” on the same train dataset. The ADS’ “plan” was dropped and was instead moved into its own separate array with 7 columns. Each column represented an insurance plan from A to G. The ‘time’ variable was also dropped in favor of its integer counterparts ‘hour’ and ‘minute’. The ADS also encoded state and car_value into integers for ease of classification.

A random forest classifier was fit on the train and its labels. The number of estimators input was 600, ‘entropy’ criterion for splitting trees which maximized information gain, and a random state value of 42.

The ADS was not validated in the code, but presumably one would use the “test_v2” submission file in order to figure out whether the predicted values fit for train match the actual values of the test dataset.

IV. Outcomes

We self validated the code by making predictions based on the test values, then checking them against the actual values of the test dataset. There was some preprocessing that had to be done before we could even use the random forest classifier to make predictions. First, we had to change the number of estimators from 600 to 100 because the runtime was too long. We also had to set the maximum depth of the tree to 5 because it kept overloading the RAM and crashing the notebook. There were 4 NaNs in the risk factor column at the start of the indices that we decided to back fill. These were all assigned a value of 4 and had negligible impact on the results. There were also 678 NaNs in the

location value. We decided to set those NaNs to the mode of location, which was “11179”. Presumably a zip code. After making predictions, we also discovered that the y_test had two columns for C, one of the insurance plan choices. We found that the two duplicate C columns were completely identical, so we removed one of them. Accuracy was calculated on an insurance plan basis. Each group had their own accuracy which can be seen in the notebook. The overall accuracy was 60%. We then decided to plot the actual and predicted values for the insurance plans. Surprisingly, the model misses a lot more than you’d expect and 60% is a number that hides a great deal of problems. Numerous models were completely missing potential values because the random forest just never even decided to spit out those values. For example, the predicted values of F,G and C are completely missing 4. A is completely missing 2. The distributions are also grossly off for all the insurance plans except E. We then decided to use the XGBoost tree in order to analyze the data. We decided to directly look at A, and why there were no values predicted to be ‘2’. Similar to the original random forest, the XGBoost tree also under-predicted the option designated by 2 in the A insurance plan, and over-predicted the values of the option designated by 1. Using the XGBoost tree, we plotted the importance of features and found that when taking into account the number of times a feature appeared in a tree, “state” was at the top of the list with “location” being fourth. We deemed these two to be protected attributes since they can be grounds for discrimination. People in NY being offered more expensive plans just because they live in NY can certainly have a disparate impact on poorer NYC residents who can’t afford such plans. After the fact, when looking at the distribution of insurance plans per state, Florida leads NYC in insurance offerings. Overall, Florida and NYC lead with the most

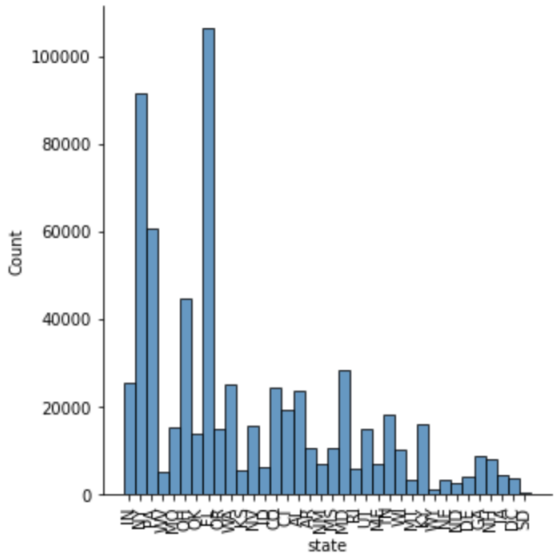
insurance options compared to any other states. In addition, there are significantly more single individuals than married couples in the dataset. This will sway the features as couples are more likely to be more established and have longer years of previous insurance than singles in the data denoted by the importance of features “C_previous” and “duration_previous”. This is further confirmed by married couples being more likely, than single couples, to be categorized as low risk. This results in single individuals having higher premiums than married couples— an incentive only for the married. Lastly, although there are fewer renters than homeowners in the dataset, the difference between the two subgroups are negligible compared to the disparity between married and single individuals in the dataset.

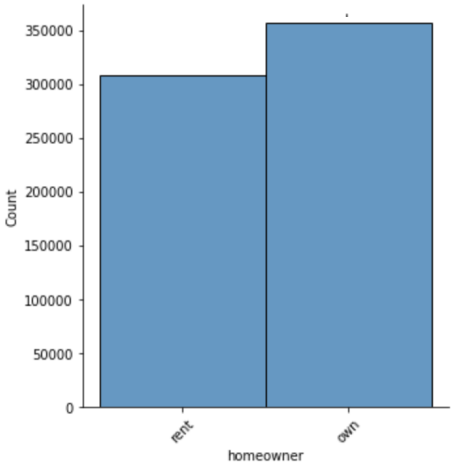
All Customers

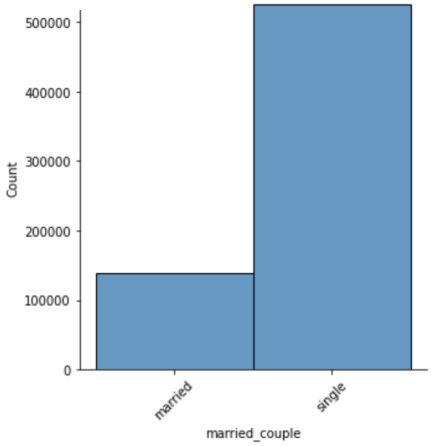
risk_factor married_couple	high	low	mid	mid-high
married	21214	27448	24434	22697
single	89540	72028	72596	94874

All Customers

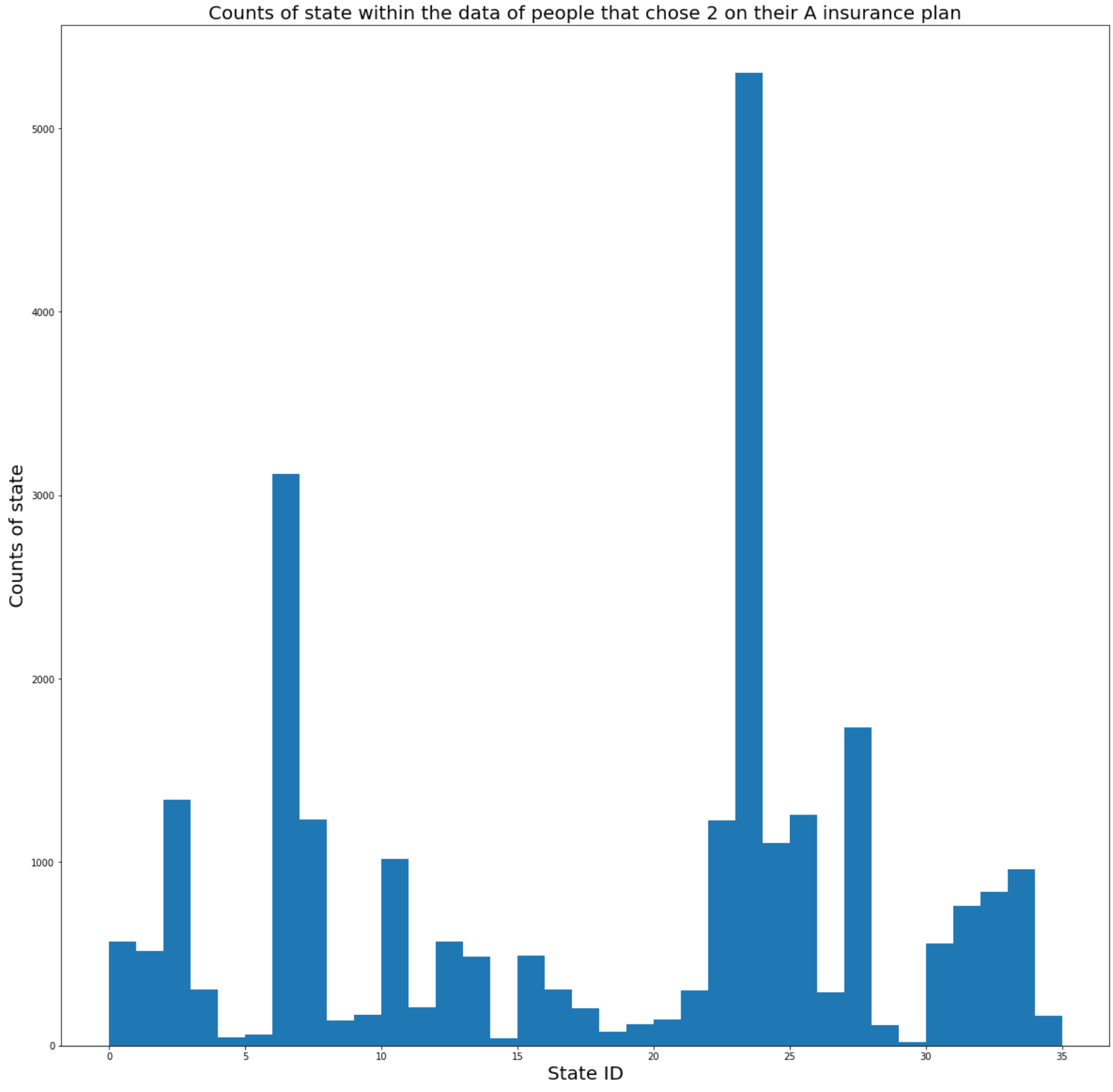
risk_factor age_youngest	high	low	mid	mid-high
21-25	16400	5906	10954	16301
25-30	18683	6866	10305	12715
30-45	30005	15155	19586	26451
45-55	16819	13200	14508	16747
55-65	13492	20355	16661	15579
< 21	4171	1845	2858	11124
>65	11184	36149	22158	18654



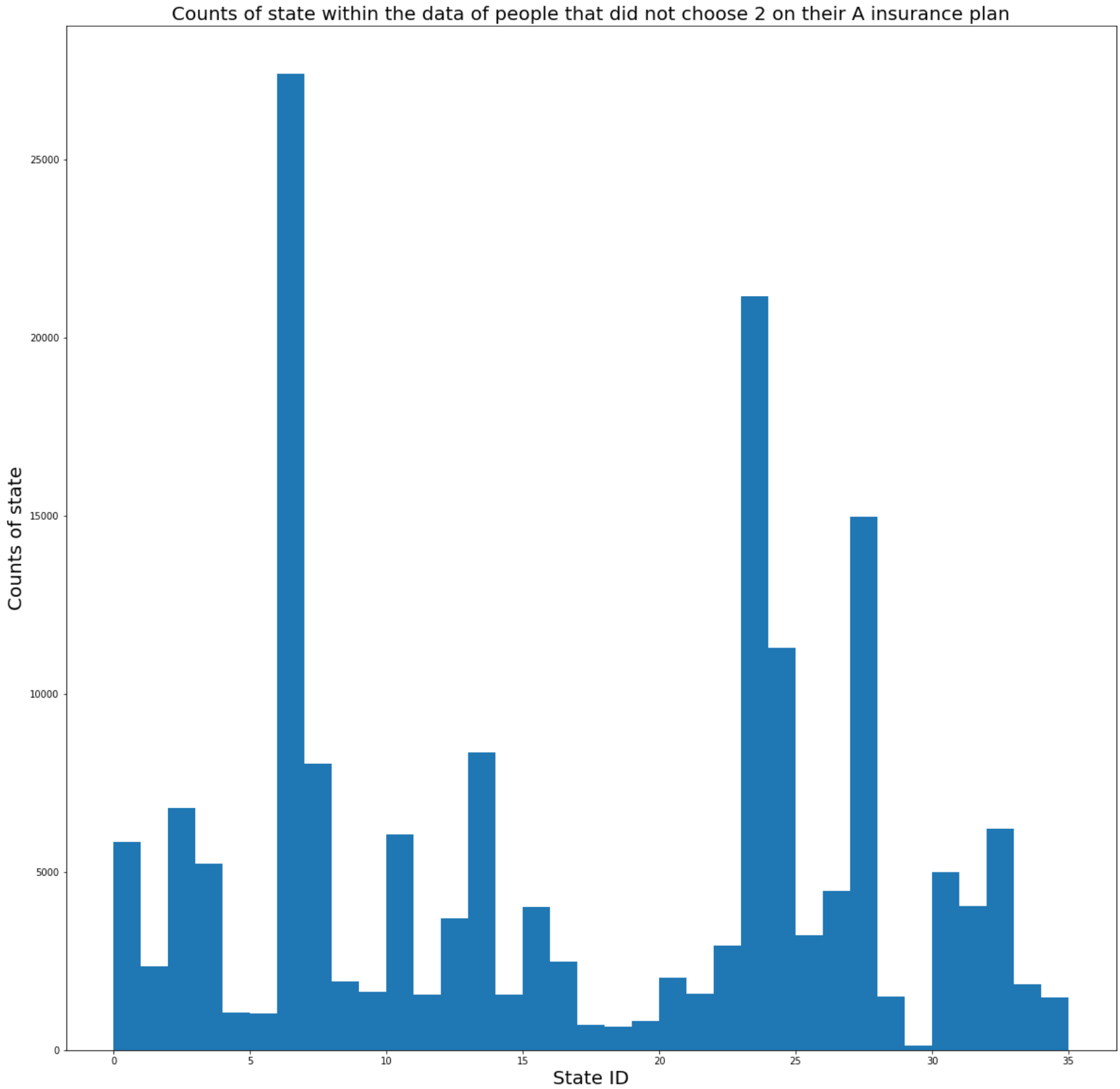




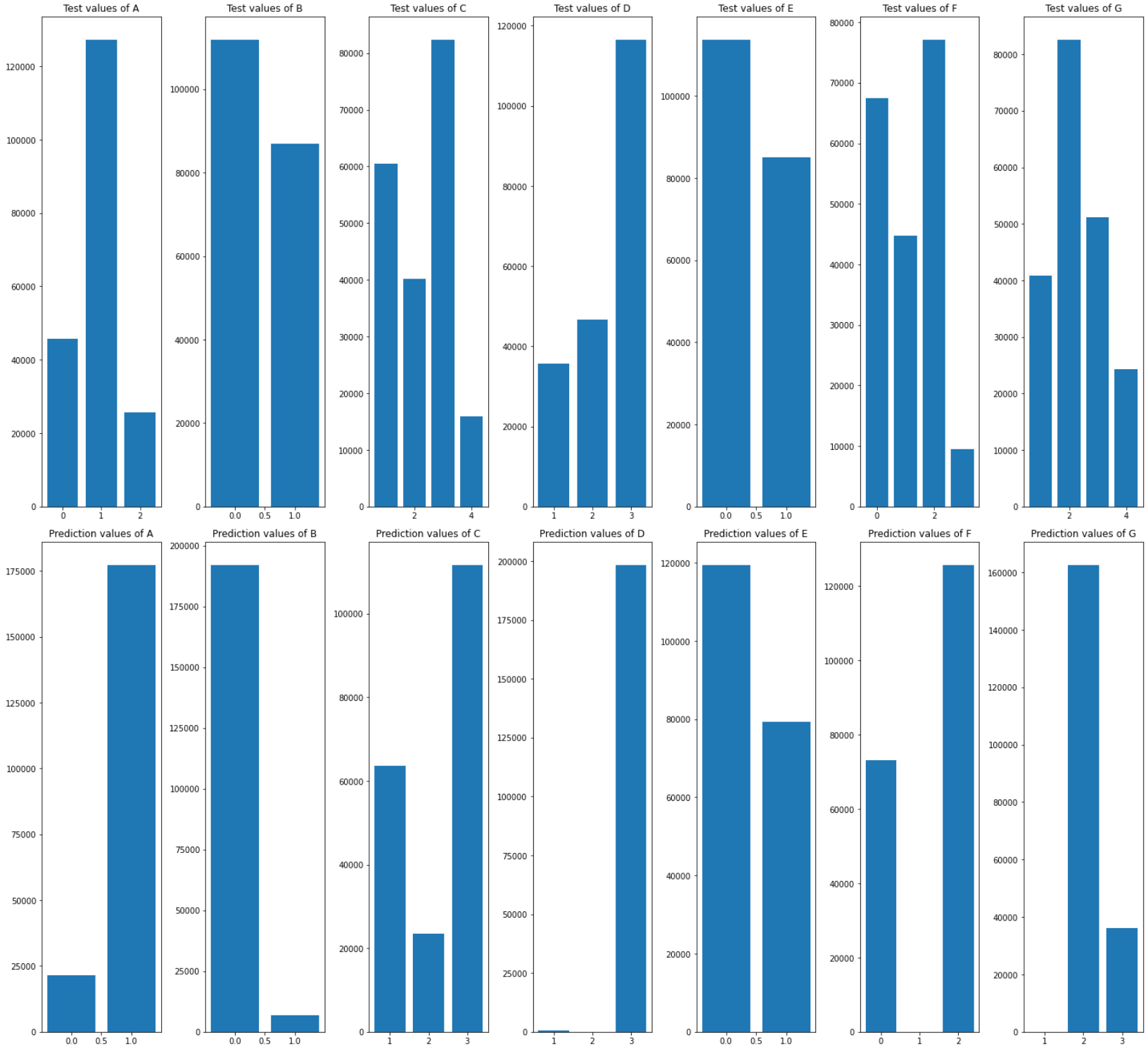
The highest count was state ID 23, which we found to be New York.

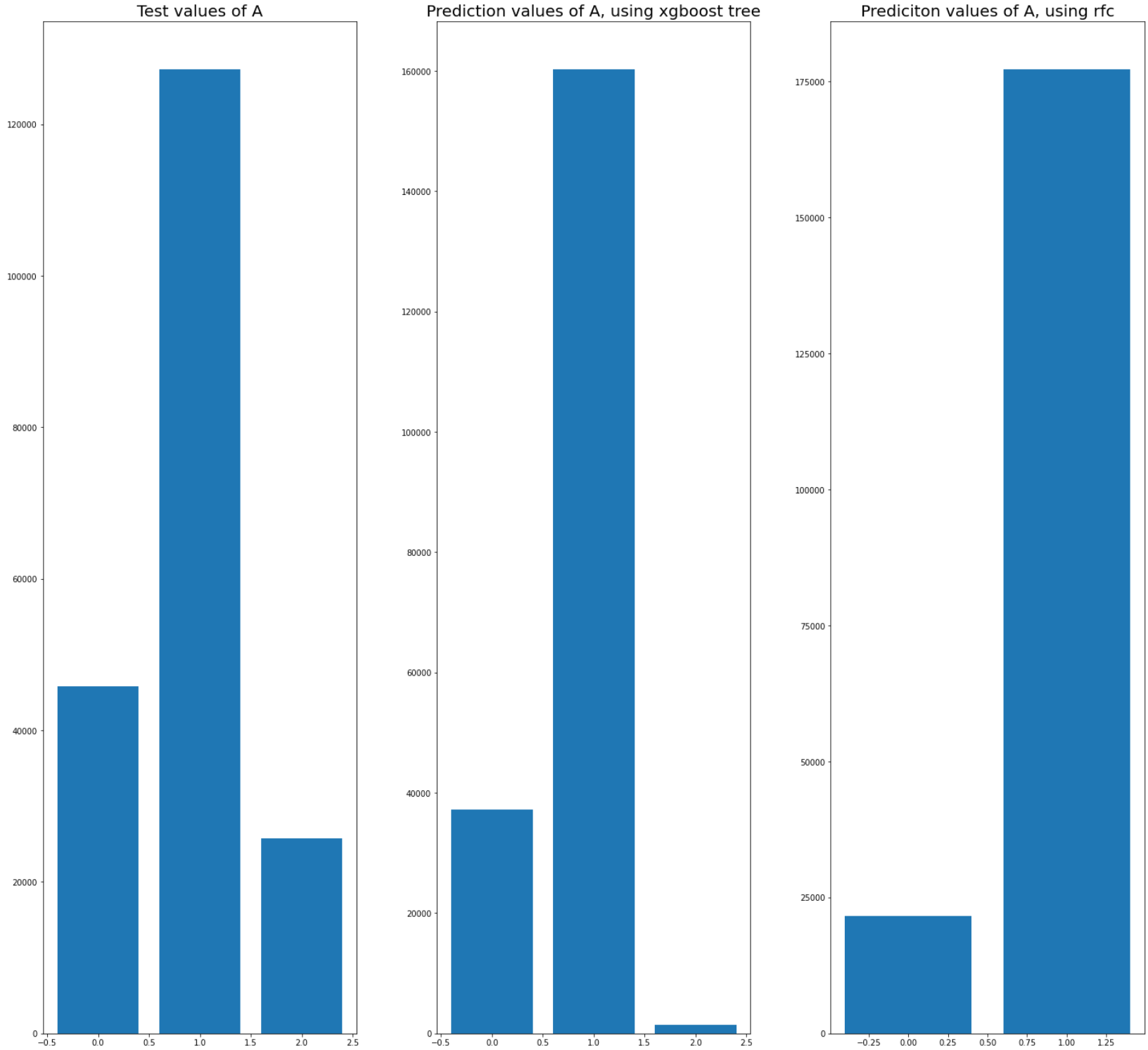


The highest count was state ID 23, which we found to be New York. There was also another value of 6 which we found to be Florida.



Counts for what type of plan was purchased for each insurance option from A to G




Counts for what type of plan was purchased for insurance option A, using xgboost tree
Test values of A
Prediction values of A, using xgboost tree
Prediciton values of A, using rfc

V. Summary

We believe that data was appropriate for this ADS; however, the ADS itself is not suitable for implementation. Although Allstate was proactive in eliminating any identifying fields or information about individuals in the data provided, an important must for protecting privacy, the ADS owner failed to exhibit care for the nuances pertaining to handling NaNs, duplicates and noise in the data. Such mishandling is evident in the lack of methods used during pre-processing prior to fitting the data on a random forest model. The data required slightly more knowledge than was specified in the metadata. We find that the implementation is extremely not robust, and the 60% overall accuracy hides a lot of problems. Such as certain values not appearing and distributions being totally off. It seems that the ADS tries to hide this by simply brute forcing the tree by setting the number of estimators to 600 which is an absurd amount. We presume that data collection was from the result of procuring information from browser cookies and was an interesting dataset to think through. Since Allstate used browser cookies and there is less common knowledge about how cookies work, more metadata or description about the “shopping_point” field should be included. We believe that Allstate did not include this description for liability purposes but, it is clear that obtaining a customers shopping history is through obtaining browser cookies; Allstate, however, does this indirectly through paying websites for the cookies their traffic brings in. The data doesn’t have sufficient metadata on what the values mean and has odd redundancies like “c_previous”.