Homework 1

Group 4

BUAN 6337 Predictive Analytics using SAS

Professor B.P.S. Murthi

Sept 4, 2020

Lin, Shih-Min

Devineni, Rajasekhar

Talreja, Vinita Govind

Zhou, Wei

Homework 1

Read the insurance claims dataset.

Using SAS answer the following questions and interpret the output in each question. In other words, **focus on what do you learn** from the output of each question.

1. What is the distribution of gender, vehicle size, and vehicle class?
2. What is the average customer lifetime value of each level of gender, vehicle size, and vehicle class?
3. Do Large cars have a higher lifetime value than medsize cars. Do a ttest and report on your findings.
4. Is there a significant difference between men and women in customer lifetime value?
5. Use ANOVA to test whether there is difference in customer lifetime value across different sales channels. Which sales channel generates the highest lifetime value?
6. What demographic factors (education, income, marital\_status) affect customer lifetime value?
7. Is there a relationship between renew\_offer\_type and response (use Chi-sq test)? Which offer type generates the highest response rate?
8. Do different renew\_offer\_types have different lifetime values? Which offer type is the best?
9. Is the effectiveness of renew\_offer\_type different across different states with respect to lifetime value?
10. What other interesting insights that are useful to the company in terms of action can be obtained from the data?
    1. Write any three (3) hypotheses. The hypotheses should be useful to the insurance firm.
    2. Do appropriate statistical tests or analysis.
    3. Report what you found in each case and also write how management can use this information to improve their operations.

Submit the following on eLearning. On the coursepage there is a link “Homework Submissions”.

One student per group should submit the following two items.

* SAS code
* Answers to questions in Word document.

Read the data

|  |
| --- |
| **Data** a1;  Infile 'D:\UTDallas\OneDrive - The University of Texas at Dallas\0Coursework\0 BUAN 6337 - PREDICTIVE ANALYTICS USING SAS\HW\car\_insurance\_19.csv' DLM=',' Missover firstobs=**2**;  Length Customer $ **12** State $ **12** Education $ **20** EmploymentStatus $ **15** Gender $ **2** Policy\_Type $ **15** Policy $ **15** Sales\_Channel $ **12** Vehicle\_Class $ **15**;  Input Customer$ State$ Customer\_Lifetime\_Value Response$ Coverage$ Education$ Effective\_To\_Date$ Employment\_Status$ Gender$ Income Location\_Code$ Marital\_Status$ Monthly\_Premium\_Auto Months\_Since\_Last\_Claim Months\_Since\_Policy\_Inception Number\_of\_Open\_Complaints Number\_of\_Policies Policy\_Type$ Policy$ Renew\_Offer\_Type$ Sales\_Channel$ Total\_Claim\_Amount Vehicle\_Class$ Vehicle\_Size$ ;  Informat Effective\_To\_Date mmddyy10.;  **Run**; |
| 9134 observations and 24 variable |

1. What is the distribution of gender, vehicle size, and vehicle class?

|  |
| --- |
| **Proc** **freq** ;  table Gender Vehicle\_Size Vehicle\_Class;  **Run**;  **Proc** **print**;  **Run**; |
|  |
| * The dataset contains approximately equal number of female and male population. * Medium sized cars are the most bought vehicles with over 70% frequency. * Four door cars are popular among the vehicle class with 50% frequency |

1. What is the average customer lifetime value of each level of gender, vehicle size, and vehicle class?

|  |
| --- |
| **Proc** **means**;  var Customer\_Lifetime\_Value; class Gender;  **Run**;  **Proc** **means**;  var Customer\_Lifetime\_Value; class Vehicle\_Size;  **Run**;  **Proc** **means**;  var Customer\_Lifetime\_Value; class Vehicle\_Class;  **Run**; |
|  |
| |  |  |  | | --- | --- | --- | | **Classes** | **Level** | **Average Customer Lifetime Value** | | **Gender** | Female | 8096.60 | |  | Male | 7909.55 | | **Vehicle size** | Large | 7545.00 | |  | Medsize | 8050.66 | |  | Small | 8085.10 | | **Vehicle class** | Four-Doo | 6631.73 | |  | Luxury C | 17053.35 | |  | Luxury S | 17123.00 | |  | SUV | 10443.51 | |  | Sports C | 750.99 | |  | Two-Door | 6671.03 | |

1. Do Large cars have a higher lifetime value than medsize cars. Do a ttest and report on your findings.

|  |
| --- |
| **Proc** **ttest**;  var Customer\_Lifetime\_Value;  class Vehicle\_Size; where Vehicle\_Size in ("Large", "Medsize");  **Run**; |
|  |
| H0: Larger cars have lifetime value less than or equal to the medium size cars.  H1: Larger cars have lifetime value greater than the medium size cars.   * Assumption: null hypothesis is true. * Since p > 0.05 in Equality of variances, it means that the variances are equal. Hence the pooled method applies. * p=0.0329 for the pooled variance. Since p < 0.05, we reject the H0. * Hence larger cars have a higher lifetime value than medium sized cars. |

1. Is there a significant difference between men and women in customer lifetime value?

|  |
| --- |
| **Proc** **ttest**;  var Customer\_Lifetime\_Value;  class Gender;  **Run**; |
|  |
| H0: No significant difference between men and women in customer lifetime value.  H1: Has a significant difference between men and women in customer lifetime value.   * Equality of variance shows that p >0.05 which means that it is a pooled variance. * Pooled variance shows a p-value > 0.05. We cannot reject the null hypothesis. * There is no significant difference in men and women in customer lifetime value. |

1. Use ANOVA to test whether there is difference in customer lifetime value across different sales channels. Which sales channel generates the highest lifetime value?

|  |
| --- |
| **Proc** **ANOVA**;  class Sales\_Channel;  model Customer\_Lifetime\_Value=Sales\_Channel;  **Run**; |
|  |
| H0: All the means of customer lifetime value of different sales channel are equal.  H1: At least two means of customer lifetime value of different sales channel differ.   * Since p > 0.05 we do not reject the null hypothesis. * Hence, all the means of customer lifetime value of different sales channel are equal. |

1. What demographic factors (education, income, marital\_status) affect customer lifetime value?

|  |
| --- |
| **DATA** A2;  SET a1;  if Marital\_Status="Single" then status=**0**;  if Marital\_Status="Married" then status=**1**;  if Marital\_Status="Divorced" then status=**2**;  if Education="High School or Below" then Education\_level=**0**;  if Education="College" then Education\_level=**1**;  if Education="Bachelor" then Education\_level=**2**;  if Education="Master" then Education\_level=**3**;  if Education="Doctor" then Education\_level=**4**;  if Employment\_status="Disabled" then Employ\_status=**0**;  if Employment\_status="Employed" then Employ\_status=**1**;  if Employment\_status="Unemployed" then Employ\_status=**2**;  if Employment\_status="Retired" then Employ\_status=**3**;  if Employment\_status="Medical Leave" then Employ\_status=**4**;  if Location\_Code="Rural" then LC=**0**;  if Location\_Code="Urban" then LC=**2**;  if Location\_Code="Suburban" then LC=**1**;  **run**;  **Proc** **REG** data=a2;  model Customer\_Lifetime\_Value=Education\_level Income status;  **Run**; |
|  |
| Considering the three demographic variables Income, Education\_Level and Marital Status. We find that Education level and Marital Status affect Customer Lifetime Value in 95% of the cases, while Income variable affects Customer Lifetime Value in 93% of the cases. |

1. Is there a relationship between renew\_offer\_type and response (use Chi-sq test)? Which offer type generates the highest response rate?

|  |
| --- |
| **Proc** **freq** ;  tables Renew\_Offer\_Type\* Response/chisq;  **Run**; |
|  |
| * There is a relationship between renew offer type and response since the calculated chi-sq value (548.1645) is greater than the table chi-sq value. * Offer 2 generates the most "yes" responses compared to other offers with a value of 23.38%. |

1. Do different renew\_offer\_types have different lifetime values? Which offer type is the best?

|  |
| --- |
| **Proc** **ANOVA**;  class Renew\_Offer\_Type;  model Customer\_Lifetime\_Value=Renew\_Offer\_Type;  means Renew\_Offer\_Type/TUKEY CLDIFF;  **Run**; |
|  |
| H0: All the means of lifetime value of different renew offer types are equal.  H1: At least two means of lifetime value of different renew offer types are differ.   * Since p < 0.05 we reject the null hypothesis. Hence, at least of two lifetime value of different renew offer types differ. * Tukey HSD test to find the best among the offers. The difference between all possible combinations of different offers shows that Offer1 is the best. |

1. Is the effectiveness of renew\_offer\_type different across different states with respect to lifetime value?

|  |
| --- |
| **Proc** **means**;  var Customer\_Lifetime\_Value; class Renew\_Offer\_Type State;  **Run**; |
|  |
| The Offer One is the most effective type across the different regions, according to the number of observations. Offer Three has the highest averaged lifetime value in Nevada. However, the Offer One provides the highest average lifetime value in the rest of the four regions.  Higher the customer lifetime value, more effective is the renew offer. Offer 1 is the most effective offer for all states except Nevada. Hence, we conclude that there is a variation in the effectiveness of renew offer type across states. |

1. What other interesting insights that are useful to the company in terms of action can be obtained from the data?
   1. Write any three (3) hypotheses. The hypotheses should be useful to the insurance firm.
   2. Do appropriate statistical tests or analysis.
   3. Report what you found in each case and also write how management can use this information to improve their operations.
2. Do different sales channel have different lifetime values? Which channel is the best?

|  |
| --- |
| **Proc** **ANOVA** plots=none;  class Sales\_Channel;  model Customer\_Lifetime\_Value=Sales\_Channel;  means Sales\_Channel/TUKEY CLDIFF;  **Run**; |
|  |
| H0: All the means of lifetime value of different sales channel are equal.  H1: At least two means of lifetime value of different sales channel are differ.   * Since p > 0.05 we do not reject the null hypothesis. * There are no significant differences between different sales channel. The insurance company can develop equal among different sales channel. |

1. Do different auto policy have different lifetime values? Which offer type is the best?

|  |
| --- |
| **Proc** **ANOVA**;  class Policy\_Type;  model Customer\_Lifetime\_Value=Policy\_Type;  means Policy\_Type/TUKEY CLDIFF;  **Run**; |
|  |
| H0: All the means of lifetime value of different auto policy are equal.  H1: At least two means of lifetime value of different auto policy are differ.   * Since p > 0.05 we do not reject the null hypothesis. * There are no significant differences between different auto policies. The insurance company can develop equal among different auto policies. |

1. Rural customers are less valuable than Urban customers

|  |
| --- |
| **Proc** **ttest**;  var Customer\_Lifetime\_Value;  class Location\_Code; where Location\_Code in ("Rural", "Urban");  **Run**; |
|  |
| H0: No significant difference between rural and urban in customer lifetime value.  H1: Has a significant difference between rural and urban in customer lifetime value.   * Equality of variance shows that p >0.05 which means that it is a pooled variance. * Pooled variance shows a p-value > 0.05. We cannot reject the null hypothesis. * There is no significant difference in rural and urban in customer lifetime value. |