Specialist Certificate in Data Analytics for Marketing

*GitHub Repository URL:* [*https://github.com/Panwar-Deepak/UCDPA\_DeepakPanwar*](https://github.com/Panwar-Deepak/UCDPA_DeepakPanwar)

1. **Introduction**

The age of digitalisation offers an abundance of information at our fingertips. As a result, it has become increasingly important for data scientists to employ the right models and methods to analyse data for various reasons. In this report, two such means of analysis are used which are A|B testing and customer churn prediction.

A|B testing is a type of experiment used to find a causal relationship and test different hypotheses. Businesses often use A|B testing to evaluate new features of products (Kaufmann et al., 2014). The common feature of A|B testing is to divide users/participants into two distinct categories: control and treatment. The treatment group is persistently subjected to a new environment whereas the control group’s environment is unchanged.

On the other hand, customer churn prediction aims to predict the likelihood of customers leaving the business by identifying early signs. This is a particularly important metric for businesses as the cost of retaining customers is generally much lower than acquiring new customers (Vafeiadis et al., 2015).

Two sets of data have been chosen from an open source for the purposes of this report. The report is thus contains two separate main sections, A|B testing and Predicting Customer Churn, which are further divided to three sub-sections each. Data analysis sub-sections aim to describe the raw data, explain the purposes of using the particular testing on the dataset as well as the necessary steps taken to prepare the data. Then, Data visualisation-Insights sub-sections present more descriptive information of the cleaned data and important insights from the visual charts. Finally, Data modelling sub-sections details the result of the actual tests, supported by other useful statistical measurements.

1. **A|B Testing**

The *Landing\_Page.csv* dataset (*Figure 1)* was taken from an open source and will be used for the A|B testing experiment. The source of the dataset is assumed to be from an ecommerce website as indicated by the 5 columns, which are *user\_id, timestamp, group, landing\_page* and *converted*. Each user was given a user identification number or *user\_id* where the following were recorded:

* *timestamp -* the time he/she visited the website,
* *group -* which group he/she was assigned under
* *landing\_page -* which page was presented,
* *converted* - whether the user performed the desired action such as buying the company’s product (converted = 1) or not (converted = 0).

If a certain user was part of the control group, the user landed on the old page (*old\_page*). For example, in *Figure 1* the user 851104 in the row 0 was assigned to the control group and was presented with the old page. On the other hand, user 661590 in the row 2 was part of the treatment group and was given the new page. Both users did not convert because the converted column shows the value 0.

Text

Description automatically generated

Figure 1. A screenshot of Landing\_Page.csv dataset

The main puzzle is whether to keep the old landing page or to replace it with a new one. As A|B testing is particularly useful to test new features in applications and products, it is a suitable way to test this puzzle.

* 1. **Data Analysis**

While doing the exploratory data analysis, no null value was found in any of the columns. However, further delving into the dataset revealed duplicate users. For instance, the same users were exposed to both new and old landing pages irrespective of which group they were assigned to. For example in *Figure 2*, the following users 630052 and 630126 were given the new and old pages at separate times. As A|B testing relies on the consistency of environment for each participant, this type of problem directly undermines the very principles of A|B testing.

Graphical user interface, application

Description automatically generated

Figure 2. Examples of duplicate users found across both treatment and control groups.

At the same time, there were duplicate records of the same users who logged into the website more than once. For example in *Figure 3*, user 773192 were recorded twice where only the *timestamp* differed. This created multitudes of the same values in the data, which can potentially create biasness in the results.

Graphical user interface, text, application, email

Description automatically generated

Figure 3. Examples of duplicate records found with same details, just logged in different times.

Therefore, duplicate users were removed from the dataset as part of the data preparation for A|B testing. Firstly, the following command was used to remove any duplicate users such as the one in *Figure* *3* by only keeping the first record of every unique user:

1. data = data.drop\_duplicates(subset='user\_id',keep='first')

Then as can be seen from *Figure 4*, any records that contained the following two combinations of *control* and *new\_page* as well as *treatment* and *old\_page* were removed using the following command.

1. data = raw\_data.loc[(raw\_data.group == 'control') & (raw\_data.landing\_page == 'old\_page')|(raw\_data.group == 'treatment') & (raw\_data.landing\_page == 'new\_page')]

A screenshot of a computer

Description automatically generated with low confidence

Figure 4. Removing users in control group who got exposed to the new page and the users in treatment group who got exposed to the old page.

* 1. **Data Visualisation - Insights**

Once the data is cleaned, a series of charts are presented below to visualise the data.

The below bar chart in *Figure 5* visualises the number of unique users in the control and treatment groups. It shows that both groups have similar number of unique users.

On the other hand, the bar chart in *Figure 6* shows the number of converted users in each group. It shows that the users in the control groups, i.e., the users who were given the old page, converted slightly more than the users in the treatment group who were exposed to the new page.

**Chart, bar chart

Description automatically generated**

Figure 5. The number of unique users in each group.

**Chart, bar chart, treemap chart

Description automatically generated**

Figure 6. The number of converted users in each group.

**Chart, pie chart

Description automatically generated**A pie chart is another great way to show proportions. The pie chart in *Figure* 7 clearly shows the proportion of the converted users in each group as compared to the users who did not convert. While a similar number of users were converted from each group, it also shows that the majority of the users in both groups did not convert.

Figure 7. The proportion of converted users in each group.

Similarly, the bar chart in *Figure* 8 shows the comparison between the converted and un-converted users, where the vertical axis is marked in 20,000s. Close to 130,000 users in each group did not convert while just about 30,000 users converted in each group.

A picture containing chart

Description automatically generated**Chart, bar chart

Description automatically generated**

Figure 9. Screenshot to show the exact count of users in each group in terms of their conversion.

Figure 8. Histogram of converted users in each group versus the users who did not convert.

*Figure* 9 shows the exact count of users in control and treatment groups in terms of their conversion. 127,785 users in control and 128,047 users in treatment group did not convert (convert=0). Whereas, 17,489 users in control and 17,264 users in treatment group converted (convert=1).

This preliminary stage alone shows that the new landing page does not perform well, and the business should likely not invest in the new landing page. However, further test is required to prove this hypothesis statistically.

* 1. **Data Modelling**

Having done an initial investigation of the data, the following hypotheses are formed. The null hypothesis (Ho) is that the control and treatment groups are same in terms of their conversion rate. On the other hand, the alternative hypothesis (Ha) is that conversion rates for the two groups will differ, i.e. the new landing page will have some impact.

Hₒ: p = pₒ

Hₐ: p ≠ pₒ

The overall conversion rate is calculated to be 11.95%. The conversion rate for the control group is 12.03% and 11.88% for the treatment group (please see *Figure 10* below).

Graphical user interface, application

Description automatically generated

Figure 10. Conversion rate, standard deviation and standard error of each group.

The results of the hypothesis testing are as follows.

|  |  |
| --- | --- |
| z\_score = 1.31 | CI 95% - treatment group: [0.117, 0.120] |
| p\_value = 0.190 | CI 95% - control group: [0.119, 0.122] |

Since our p-value is above the α=0.05 threshold and the z value is 1.312 < 1.96, the null hypothesis, Hₒ, cannot be rejected (Failed to reject Null Hypothesis). In other words, both the groups are similar in terms of conversion rate.

Furthermore, we can say with 95% confidence that the treatment conversion rate lies between 11.7% - 12%. Similarly, with the same confidence of 95%, we can say that the control conversion rate lies between 11.9% - 12.2% which is slightly better. This is a further proof that the new design is not likely to be an improvement on the old design.

In conclusion, from the business point of view, the new landing page does not make a convincing case for investment.

1. **Predicting Customer Churn**

The Churn\_Prediction.csv dataset (*Figure* 11) is the second dataset used in this report. It was also taken from an open source and will be used to predict customer churn. The dataset is presumed to be from a bank as suggested by the data columns such as *Balance* which likely refers to bank balance and *HasCrCard* to customers who have credit cards. The customer churn prediction aims to identify which customers are likely to be leaving the bank. There are 14 variables in total. The variable *exited* is the dependent variable while the other variables are independent.

A screenshot of a computer

Description automatically generated with medium confidence

Figure 11. A screenshot of Church\_Prediction.csv dataset.

* 1. **Data Analysis**

No null value was found inside the dataset which was determined using the following function (iii).

1. raw\_data.isnull().sum()

The above function (iv) was then used from Pandas to understand the data present in each column as shown in *Figure* 12. For example, we can see that the average age of customers is 38.92 while the average credit score is 650.53.

1. df.describe()

Graphical user interface, text

Description automatically generated

Figure 12. Overview of the data using descriptive statistics.

As for the dependent variable, exited, 7963 customers remained (exited=0) while 2037 customers left (exited=1).

The following code (v) was used to calculate the number of unique values in each column (*Figure* 13) where it revealed the maximum number of categories as 11.

1. for col in raw\_data.columns:

print("{}: {}".format(col,raw\_data[col].nunique()))

Text

Description automatically generated

Figure 13. Number of unique values in each column

Using this as one of the conditions, an user defined function called *col\_name* was created which distinguished categorical or numerical variables based on user requirements.

Timeline

Description automatically generated

Figure 14. Correlation between independent and dependent variables.

The correlation between variables were not strong as shown in *Figure* 14. Age and balance showed the strongest correlation compared to other variables which were only 0.29 and 0.12.

* 1. **Data Visualisation - Insights**

Once the initial investigation of the data is complete, series of visual charts are created to present the data. Firstly, user defined function plots which take data frames and column names as input were created below in *Figure* 15.

Chart, bar chart

Description automatically generated

Figure 15. User defined function plot.

A lot of useful information can be seen from the above plots. For example, the second plot in the first row shows that female customers left more than men, while the German customers left more than French and Spanish customers. Also, customers with credit card were more likely to leave and customers who were not signed up to more products were more likely to leave compared to those who were.

Chart, histogram

Description automatically generated

Figure 16. Age distribution of leavers (exited=1) and not leavers (exited=0).

The above graph in *Figure* 16 shows the age distribution between customers who left versus those who did not. The graph for the not leavers peaks earlier than the one for leavers, meaning older customers between 40-50 were more likely to leave whereas customers between the ages of 30-40 remained. The bank should focus on retaining customers within the age bracket of 40-50.

Chart, histogram

Description automatically generated

Figure 17. Age distribution of users (both leavers and not leavers).

The distribution graph in *Figure* 17 shows that customer’s exited value climbed from 35, peaking at 55. Then the graph falls sharply. It means that older customers are more likely to leave until the age of 55 and more likely to remain henceforth.

Chart, histogram

Description automatically generated

Figure 17. Credit score distribution of leavers (exited=1) and not leavers (exited=0).

As for credit score distribution in *Figure* 18, both distribution is almost symmetrical with majority of the data concentrating in the middle. However, the peak or the mode of the not leavers is located further to the right around 650-700 than the leavers’ which is around 600-650. It means the customers who remained tend to have higher credit score while customer with lower credit score were likely to leave. The bank should focus on retaining customers with lower credit score.

Chart, histogram

Description automatically generated

Figure 18. Balance distribution of leavers (exited=1) and not leavers (exited=0).

The balance distribution between leavers and non leavers were very similar in *Figure* 18. It means the amount of bank balance did not affect the customers’ decisions to leave the bank majorly.

* 1. **Data Modelling**

As it is a binary classification problem, logistic regression and random forests were chosen. The test accuracy for logistic regression was 84.24% whereas it was 86.3% for random forest (*Figure 19)*. Random forest model was tuned by using the following command (vi) to achieve 86.35% accuracy.

1. model\_cv\_rf.best\_params\_

Chart, line chart

Description automatically generated

Figure 19. ROC-AUC

In conclusion, random forest performed better than logistic regression model to predict customer churn for this problem.

## References

* Vafeiadis, T., Diamantaras, K.I., Sarigiannidis, G., Chatzisavvas, K.Ch. (2015). A comparison of machine learning techniques for customer churn prediction. *Simulation Modelling Practice and Theory*, *55*, 1-9. <https://doi.org/10.1016/j.simpat.2015.03.003>.
* Kaufmann, E., Cappé, O., & Garivier, A. (2014, May). On the complexity of A/B testing. *Conference on Learning Theory,* *35*(1–23), 461-481. PMLR.