|  |  |
| --- | --- |
| How to Travel Around the World | Eagle Creek  **INCOME TRAVEL INSURANCE PORTFOLIO ANALYSIS PROJECT 2025**  *Summer Internship May-Aug 2025* | Abstract  This is an innovation project started in May 2025 by Actuarial team to discover use cases from travel insurance per-trip policy analysis on the grounds of risk behaviour and purchasing behaviour of customers.  Bervyn Wong  GI Innovation & Analytics Intern, AT Dept |

**General Introduction**

Income Insurance, founded in 1970, is a local co-operative that provides life, health, motor, travel and other general insurance products. It is currently the largest composite insurer in Singapore, as well as one of the largest general and health insurance providers in the market. It is also the largest motor insurer in Singapore, covering close to 25% of vehicles.

**Context**

The Singapore Travel Insurance Market size was valued at USD143.2 million in 2023 and is predicted to reach USD646.7 million by 2030, with a CAGR of 22.6% from 2024 to 2030. The proliferation of online platforms for purchasing travel insurance has significantly transformed the landscape of the travel insurance industry within the country, improving accessibility and convenience for consumers. Furthermore, the online presence of these platforms streamlines the purchasing process, eliminating the need for lengthy paperwork or visits to insurance providers in person. As a result, the increasing prevalence of online platforms has intensified competition within the industry, compelling insurance companies to innovate and offer more competitive rates and comprehensive coverage to attract customers in the digital age.

**Overview of Income's travel insurance products**

Within Income's portfolio of travel insurance products, are offered several tiers of plans, namely 'Classic', 'Deluxe' and 'Preferred', with classic being the most affordable and Preferred the most premium. Quotes of plans are given based on factors such as travel region, age, trip duration, advance purchase, customer tenure, discounts, etc. Travel products usually cover aspects like trip cancellation, loss/damage to personal belongings, personal accident and local/overseas medical expenses. Majority of travel insurance purchases take place online. With the multitude of data volume that accumulates over time, Income leverages on its ability to innovate and seek new use cases to both preserve their loyal customer base as well as acquire new and high-value customers to boost sales revenue year-on-year.

**Project overview – Introducing the concept of repeat purchasing behaviour**

The goal of the project is to find out use cases for customer-targeted promotion strategies with the help of machine learning and predictive modelling.

The project is best split into two aspects: understanding customer risk behaviour (profitability model), and purchasing behaviour (repeat purchase model).

There are currently 6 models being constructed:

* **Model 1: Will a customer make at least one additional purchase in the next 12 months?**
* Model 2: How many policies will a customer purchase in the next 12 months?
* Model 3a: How many Classic policies will a customer purchase in the next 12 months?
* Model 3b: How many Deluxe policies will a customer purchase in the next 12 months?
* Model 3c: How many Preferred policies will a customer purchase in the next 12 months?
* Model 4a: How many trips will a customer take to ASEAN in the next 12 months?
* Model 4b: How many trips will a customer take to ASIA in the next 12 months?
* Model 4c: How many trips will a customer take to WORLDWIDE in the next 12 months?
* Model 5: How last-minute will a customer be in making their next purchase?
* Model 6: What will be the total premium value a customer generates in the next 12 months?

This report focuses mainly on the latter (repeat purchase model), where we seek to identify who exactly are Income's repeat purchasers, and work towards analytical views on how we can best perform segmentation on customers who bring high value to Income.

As for a high-level overview of the end-to-end process, it includes a vigorous amount of:

* Data collection
* Data preprocessing
* Feature engineering
* Exploratory data analysis
* Model training
* Model selection
* Model deployment
* Post analysis and insights from findings

**Repeat Purchase Model: Definitions**

**1. Repeat Purchase Definition**

Repeat Purchase is defined as a customer making a subsequent travel insurance policy purchase within 1 year of their first purchase in a defined analysis period. The analysis covers two periods:

• **Period 1:** June 1, 2022 - May 30, 2023

• **Period 2:** June 1, 2023 - May 30, 2024

For each customer's first purchase within a period, the system tracks whether they purchase again within the 1-year, creating a Boolean target variable

(repeat\_purchase\_final: 1 = repeat purchaser, 0 = one-time purchaser).

**2. Feature Definitions**

**Raw Features (Original Dataset)**

• **InsuredNric:** Customer unique identifier

• **Insured\_Premium:** Policy premium amount

• **Claim\_count:** Number of claims made

• **Discounts:** Discount amount applied

• **Coverage\_final:** Coverage level (Classic/Deluxe/Preferred)

• **Region:** Geographic coverage (ASEAN/ASIA/WORLDWIDE)

• **plantype\_fgi:** Plan type (Individual/Group/Family)

• **trip\_duration:** Length of trip in days

• **Advance\_purchase:** Days between purchase and trip start

• **Tenure\_Years:** Customer relationship duration

• **Insured\_age:** Age of insured person

• **Policyholder\_age:** Age of policyholder

• **No\_of\_insured:** Number of people covered

**Engineered Features**

**Pre-Period Features (Historical Behaviour)**

Analyse customer behaviour in the 1 year before their first purchase in the analysis period:

• **Categorical Counts:** Region\_type\_pre\_final, plantype\_fgi\_pre\_final, Coverage\_type\_pre\_final - Historical preferences by category

• **Numerical Averages:** [feature]\_avg\_pre\_final - Average values for claims, premiums, trip duration, discounts, advance purchase, tenure, and claim processing days

• **Policy Count:** policies\_pre\_final - Number of policies purchased historically

• **Family Flag:** family\_flag\_pre\_final - Whether customer used family coverage historically

**Post-Period Features (Future Behavior)**

Analyse customer behaviour in the 1 year after their first purchase in the analysis period:

• **Categorical Counts:** Region\_type\_post\_final, plantype\_fgi\_post\_final, Coverage\_type\_post\_final - Future preferences by category

• **Numerical Averages:** [feature]\_avg\_post\_final - Average values for future purchases

• **Policy Count:** policies\_post\_final - Number of policies purchased subsequently

• **Family Flag:** family\_flag\_post\_final - Whether customer used family coverage in future

**Additional Features**

• **analysis\_period:** Indicates whether the customer's first purchase occurred in Period1 or Period2

• **first\_purchase\_date\_period1/period2:** Exact dates of first purchases in each period

**3. Model Structure**

The final dataset contains one row per customer per analysis period, combining:

• Original transaction data from the customer's first purchase

• Historical and future behavioural patterns (pre-period + post-period features)

• Boolean repeat purchase outcome (target variable)

This structure enables prediction of repeat purchase likelihood based on both historical customer behaviour and the characteristics of their current (first) purchase.

***\*For a more detailed description of the features, refer to features\_list.xlsx.***

***\*For a more technical documentation of the model pipeline script, refer to README.md***

**Approach**

This dataset includes ~6 million rows of travel insurance policy information. The raw dataset is currently prepared at a policy-level and is established to be the input data that feature engineering will be performed on.

**A) Data preprocessing & Feature engineering via Python**

* The input data is first passed into a script that will perform feature engineering to extract relevant features for model training
* Input dataset: "Q:\GI\Projects\2025\Travel Dynamic Pricing\Data\masterdataset\repeat\_purchase\_model\raw\_dataset\travelpolicies\_2013\_2025may.csv"
* The script implements a repeat purchase analysis pipeline for travel insurance customers
* It identifies customer's first purchases within two defined annual periods (June 2022 -May 2023 and June 2023 - May 2024), then creates comprehensive sets of intermediate by analysing customer behaviour in 1-year windows before and after these first purchases
* The pipeline then extracts three types of features: categorical (regions, plan types, coverage), numerical (claims, premiums, premium amounts, trip details), and derived metrics (policy counts, repeat purchase flags)
* It handles temporal relationships by creating lookback features (pre-purchase behaviour) and lookahead features (post-purchase outcomes), then consolidates these into period-neutral features which will be used in the modelling process
* The end goal is producing a customer-level dataset where each row represents a customer's first purchase in either period, with their historical behaviour patterns (training data) and subsequent purchase outcomes (validation data)

**B) Exploratory data analysis via Tableau**

* Input Data: "Q:\GI\Projects\2025\Travel Dynamic Pricing\Data\masterdataset\repeat\_purchase\_model\Repeat Purchase Pipeline\Model Training\travelpolicies\_2013\_2025may\_repeatpur\_finalfeatures.csv"
* Data analysis is then conducted on Tableau to extract insights and trends within the data and features selected by the model
* Since the features created are very cardinal (many unique values), and the goal is clustering customers into identifiable segments, the final predictive features are further transformed within Tableau into categorical groups/bins (refer to Appendix 'features\_list' for the full list of EDA transformed features)
* Statistical analysis tools such as Cramer's V correlation matrix is constructed to identify the highly correlated pairs of features that will allow further analysis of customer segmentation (refer to READMETOO.md for script documentation)
* It also serves to not only confirm certain hypotheses that have been made with regards to the data available, but also help give the AT team a better overview of who contributes to Income's repeat customer base through identifying highly correlated features and forming relationships among them to craft a more specific persona of customers

**C) Machine Learning model training via DataRobot**

* Input data: "Q:\GI\Projects\2025\Travel Dynamic Pricing\Data\masterdataset\repeat\_purchase\_model\Repeat Purchase Pipeline\Model Training\travelpolicies\_2013\_2025may\_repeatpur\_finalfeatures.csv"
* Once the python script has been generated with all the final features, it is passed into DataRobot for model training
* Perform feature selection and set target variable, and depending on the type of feature, it will usually either be a classification or a regression problem.
* As of the current state of the project, the focus has been primarily on only one model, which asks the question: "Will a customer make a repeat purchase in the next 1 year?"
* The effects of each predictive feature will be assessed based on how impactful it is and filtered out for further exploratory data analysis

**Insights**

A screenshot of a computer

AI-generated content may be incorrect.

Figure 1 – Repeat Purchase Correlation Matrix Analysis

Out of the final list of features, 9 features have been further transformed via identify deeper associations\* between customer behavioural variables, which is used to support feature selection decisions and customer segmentation insights for data visualisation in Tableau. The features are:

* ‘Analysis Period’
* ‘Customer Tenure Groups’
* ‘Policyholder Age Groups’
* ‘Plan Type Preference’
* ‘Coverage Preference’
* ‘Region Preference’
* ‘Advance purchase’
* ‘Trip Duration Avg Pre Final Groups’
* ‘Policies\_Per\_Year’

For a more in-depth explanation on the logic behind the further data transformation of these features, refer to ‘feature\_list.xlsx’.

\*Note that this is NOT a cause-and-effect analysis

From the findings, we have shortlisted the 10 strongest associated pairs of features ranked according to their correlation coefficients.

1. **Trip Duration x Region (0.54)** – longer trips associated with further travel regions
2. **Advance Purchase x Plan Type (0.46)** – Earlier advance purchases associated with group/family plans
3. **Customer Tenure x Plan Type (0.46)** – Shorter tenured customers associated with individual travellers
4. **Coverage x Plan Type (0.39)** – Classic plans associated with individual travellers
5. **Coverage x Customer Tenure (0.37)** – Higher coverages associated with longer customer tenures
6. **Advance Purchase x Coverage (0.36)** – Earlier advance purchase associated with higher coverages
7. **Advance Purchase x Region (0.36)** – Earlier advance purchase associated with further travel region
8. **Advance Purchase x Customer Tenure (0.36)** - Earlier advance purchase associated with longer tenures
9. **Trip Duration x Plan Type (0.34)** – Further travel regions associated with more group/family plans
10. **Region x Plan Type/Coverage (0.33)** – Further travel regions associated with more group/family plans + higher coverages

These associations are merely a starting point for multi-dimensional analysis to create a more specific and targeted customer segment and can help us to understand better who exactly our repeat purchasers are.

The next step is to perform multi-class data analysis using scatter plots to potentially identify specific clusters of customer segments. From the 7-8 features that were ranked based off the correlation matrix, below are Tableau plots that attempt to illustrate 3 statistically significant relationships.

A computer screen with a graph on it

AI-generated content may be incorrect.

Figure 2 - Relationship 1 – Primary relationship between Advance Purchase and Trip Duration, and identifying specific segments with Region, Coverage, Plan Type, PH age

From relationship 1, we observe a **moderate positive correlation** between **Advance Purchase and Trip Duration**, which depicts that customers booking further ahead take longer trips. A dense concentration in short advance/short trip pairs is observed. Similarly, **the shorter the duration of trip, the closer the region of travel**, as seen from the ASEAN (blue) dots dominating spontaneous, short-trip segments. Conversely, WORLDWIDE (Red) dots are mostly scattered, which represents diverse planning behaviours for international travel. We also observe that **Preferred coverage customers plan furthest ahead**, while **Classic coverage customers are more spontaneous in purchasing travel policies**.

From here, there is a clear understanding of market segmentation between **spontaneous regional travellers** and **strategic international planners**.

A computer screen with a graph on it

AI-generated content may be incorrect.

Figure 3 - Relationship 2 - Primary relationship between Trip Duration vs Tenure, and identifying specific segments with Region, Coverage, Plan Type, PH age

From relationship 2, we observe a **relatively weak correlation**, which suggests that **customer tenure is not truly associated with trip duration**, as seen from the horizontal concentrations around 10–30-day trips across all tenure levels. However, inferring multi-dimensional insights, we see that regarding customer evolution, there is a clear **progression between plan types**, i.e. **Classic -> Deluxe -> Preferred** as tenure increases. This suggests that **Income's longer tenured customers tend to buy from more premium plans** types.

Moreover, **ASEAN region**, i.e. shorter regions, cheaper plans usually dominate new customer segments, as defined specifically from have 0–2 years of tenure, purchasing classic coverage. **WORLDWIDE region**, i.e. customers who travel to further regions, more expensive plans contribute to higher retention, scattered across longer tenure periods.

Another valuable observation and inference are that **2-4 year tenure shows the highest Preferred coverage concentration**, which explains the optimal upselling window for such groups of customers.

A computer screen with a graph on it

AI-generated content may be incorrect.

Figure 4 - Relationship 3 - Primary relationship between Advance Purchase vs Tenure, and identifying specific segments with Region, Coverage, Plan Type, PH age

From relationship 3, we observe a **strong positive correlation. Longer tenured customer book further in advance**. There is a clear distinction between **spontaneous new customers** and **more strategic 'experienced' customers**.

Through the lens of customer tenure, we also observe the maturation of customers in their journey with Income. 0-1 years tenure mainly comprise of ASEAN, Classic travellers who are also spontaneous purchasers. 2-3 years tenure comprise of mixed regions travellers, purchasing upgraded coverages, with better trip planning. 4+ years of tenure show mostly premium coverage preferences, with strategic advance booking.

WORLDWIDE customers also show the highest advanced purchase rates across all tenure levels.

Another customer development pathway is also very clearly observed, **where higher premium coverage is directly correlated with more strategic planning behaviour**.

In summary, there are 3 clear segments this analysis helps us to identify:

**Segment 1: "Spontaneous Regional Travellers"**

• ASEAN, Classic coverage, 0-2 years tenure, short advance booking

**Segment 2: "Developing Planners"**

• Mixed ASIA/ASEAN, Deluxe coverage, 2-4 years tenure

**Segment 3: "Strategic Premium Travellers" (Highest value)**

• WORLDWIDE/ASIA, Preferred coverage, 3+ years tenure, high advance purchase

We then proceed to the next phase, which is model training.

From the model taining process on DataRobot, the image below shows the list of predictive lookback features that were used to predict the repeat\_purchase\_final target variable:

**A computer screen with a black and white background

AI-generated content may be incorrect.**

Figure 5 - Selected predictive features used for model training

The top 5 features and their individual effects on predicting the target variable are illustrated in the following snapshots.

**A computer screen with a graph on it

AI-generated content may be incorrect.**

Figure 6 - Policies Per Year vs Repeat Purchase Probability - This graph illustrates that the greater the number of policies a customer has purchased historically on average, the higher the probability of them making a repeat purchase in the next 1 year

**A computer screen with a graph on it

AI-generated content may be incorrect.**

Figure - Customer Tenure vs Repeat Purchase Probability - This graph illustrates that the longer the tenure of a customer with Income, the higher the probability of them making a repeat purchase in the next 1 year

**A computer screen with a graph on it

AI-generated content may be incorrect.**

Figure - Policyholder age vs Repeat Purchase Probability - This graph illustrates the older and more stable a customer is, the more likely they are to come back and do a repeat purchase

**A computer screen with a graph on it

AI-generated content may be incorrect.**

Figure - Trip Duration vs Repeat Purchase Probability - This graph illustrates that customer who take longer trips are more likely to come back to do a repeat purchase in the next 1 year

**A screen shot of a computer

AI-generated content may be incorrect.**

Figure - Advance Purchase vs Repeat Purchase Probability - This graph illustrates that customers who purchase earlier in advance are more likely to come back and do a repeat purchase in the next 1 year

**Conclusion and Future Directions**

**Project Summary**

This repeat purchase model pipeline has established a data foundation for understanding customer behaviour patterns across two distinct time periods (June 2022-May 2023 and June 2023-May 2024). Through feature engineering, we have transformed raw data into meaningful behavioural indicators, creating both historical (pre-period) and predictive (post-period) features that capture customer preferences, temporal purchasing patterns, and engagement levels. This understanding is also further deepened through exploratory data analysis, where correlation analysis revealed 7-8 features with relatively strong associations, providing clear direction for feature selection and model development.

As for preliminary model training and validation, the initial modelling experiments using DataRobot have provided baseline performance metrics and identified promising algorithmic approaches for the multi-class classification challenge ahead.

**Recommended Solutions and Future Directions**

**High-Level Approach**

The team is currently looking to develop a segmentation approach that goes beyond simple yes/no predictions for repeat purchases. Instead of just indicating whether a customer will come back to repeat a purchase, the new model creates three customer tiers based on likelihood scores. This gives us much more actionable insights and lets us tailor our strategies to different customer types.

**Why This Matters for Business**

Traditional models deliver binary answers, i.e. a customer will either make or not make another purchase. Multi-class segmentation goes a level further into understanding the scale of customer behaviour.

This approach segments customers into three clear groups based on their repeat purchase likelihood. This means we can focus our premium services on high-potential customers while using cost-effective strategies for lower-potential ones. For travel insurance, this directly translates to better resource allocation and potentially higher returns.

**How It Works – Modelling Process**

There are 4 steps to the process.

1. Engineer features that capture customer behaviour before and after purchases, including their preferences and patterns.
2. Generate likelihood scores using cross-validation
3. Establish percentile groups as thresholds and convert these scores into the three customer segments based on the percentile group it falls under
4. Analyse each segment separately to understand any patterns within and develop targeted strategies.

We start with prediction scores from the recently trained machine learning model with target variable **repeat\_purchase\_final**. The cross-validation prediction scores range from 0 to 1 and represent each customer's likelihood of making another purchase. Then create percentiles to create three distinct segments:

• **Segment 1:** 0-33rd percentile - Low probability repeat customers (Range: 0.1746 – 0.3713)

• **Segment 2:** 33rd-66th percentile - **Medium** probability repeat customers (Range: 0.3713 - 0.5293)

• **Segment 3:** Top third (66th-100th percentile) - High probability repeat customers (Range: 0.5293 – 0.9756)

This percentile approach automatically adapts to the existing data and ensures each segment has roughly the same number of customers, making comparisons fair and meaningful.

**Model Training on DataRobot - Using Machine Learning for Deeper Insights**

Once the segments are established, further analysis can be done to identify appropriate clustering techniques. K-means clustering within each segment helps us find sub-groups with similar behaviours. For example, within our high-likelihood customers, we might discover distinct groups like "frequent short-trip travellers" and "annual luxury travellers."

We also run comparative analyses between segments 1 and 3, i.e. the extreme cases to understand what the distinct features are separating our segments. The observations are captured under the 'Clusters' tab within 'features\_list.xlsx' appendix.

**Observations**

Within segments 1 and 3 (low vs high probability repeat purchase), 3 clusters were formed by the model. The cluster sizes vary between each segment, and below are the screenshot tables depicting a distinct persona for each segment.

A screenshot of a computer

AI-generated content may be incorrect.

From segment 3, we see that for each cluster:

* + Cluster 1: Solo travellers In their 30s/40s, who are new to Income, travel mainly to ASEAN regions, make short trips, regular buyers, purchase cheapest plans/coverage
  + Cluster 2: Solo travellers in their 30s/40s, who are new to Income, travel mainly to ASEAN regions, make short trips, occasional buyers, purchase cheapest plans/coverage
  + Cluster 3: Group travellers in their 40-60s, experienced customer with Income, travel to further regions, make medium trips, frequent buyers, purchase more premium plans/coverage

**~ END OF REPORT ~**