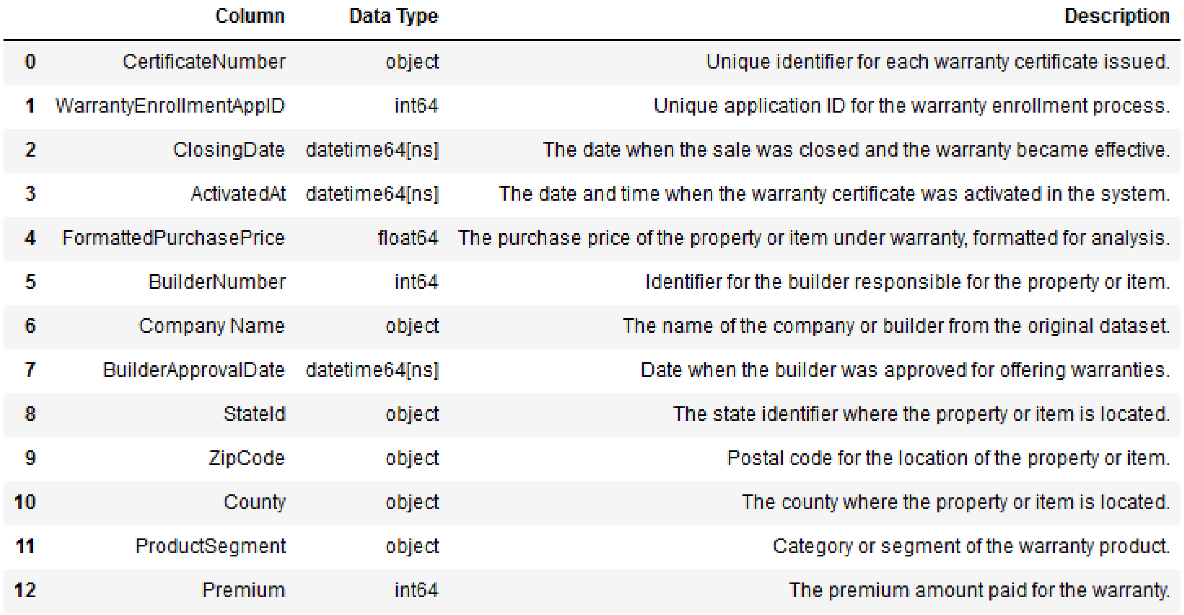
Melissa Young
DSC680 – T301 Applied Data Science
1.2 Project 1: Draft White Paper/Milestone 2 - Appendix

Appendix:

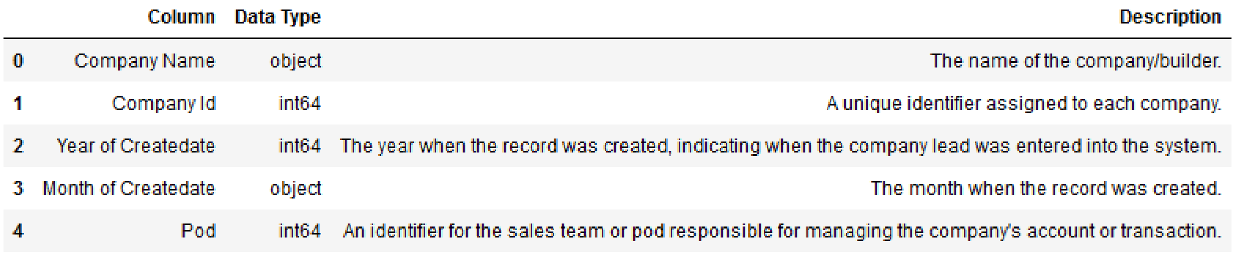
A. Data Dictionaries

A.1 CertSummTBL Data Dictionary



	Column	Data Type	Description
0	CertificateNumber	object	Unique identifier for each warranty certificate issued.
1	WarrantyEnrollmentAppID	int64	Unique application ID for the warranty enrollment process.
2	ClosingDate	datetime64[ns]	The date when the sale was closed and the warranty became effective.
3	ActivatedAt	datetime64[ns]	The date and time when the warranty certificate was activated in the system.
4	FormattedPurchasePrice	float64	The purchase price of the property or item under warranty, formatted for analysis.
5	BuilderNumber	int64	Identifier for the builder responsible for the property or item.
6	Company Name	object	The name of the company or builder from the original dataset.
7	BuilderApprovalDate	datetime64[ns]	Date when the builder was approved for offering warranties.
8	StateId	object	The state identifier where the property or item is located.
9	ZipCode	object	Postal code for the location of the property or item.
10	County	object	The county where the property or item is located.
11	ProductSegment	object	Category or segment of the warranty product.
12	Premium	int64	The premium amount paid for the warranty.

A.2 HubSpot Data Dictionary



	Column	Data Type	Description
0	Company Name	object	The name of the company/builder.
1	Company Id	int64	A unique identifier assigned to each company.
2	Year of Createdate	int64	The year when the record was created, indicating when the company lead was entered into the system.
3	Month of Createdate	object	The month when the record was created.
4	Pod	int64	An identifier for the sales team or pod responsible for managing the company's account or transaction.

B. Statistical Analyses and Model Performance

B.1 Total CRM Leads and Certificate Entries

Total CRM Leads: 3062
Total Certificate Entries: 40191

B.2 Conversion Analysis

Count of Non-Converting Leads: 2327
Count of Converting Leads: 190
Percent of Converting Leads: 8.17%

B.3 Model Performance Summary

```
Adjusted Classification Report:
              precision    recall  f1-score   support

           0       0.05      0.81      0.10        16
           1       1.00      0.81      0.90      1276

    accuracy                           0.81      1292
   macro avg       0.52      0.81      0.50      1292
weighted avg       0.99      0.81      0.89      1292

Adjusted Confusion Matrix:
 [[  13    3]
 [ 238 1038]]
```

C. Model Insights and Implications

C.1 Model Overview
The predictive model developed as part of this project is designed to enhance the strategic decision-making process within the builder warranty sector by identifying sales leads most likely to convert into actual sales. Utilizing historical data extracted from the Azure SQL Certificate Database and the HubSpot Sales CRM, the model employs a sophisticated blend of logistic regression and machine learning techniques to analyze patterns and predict outcomes.

Key Features:

Data Utilization: The model integrates a wide range of data points, including builder profiles, geographical locations, warranty details, and interaction histories from the CRM, creating a holistic view of each lead.

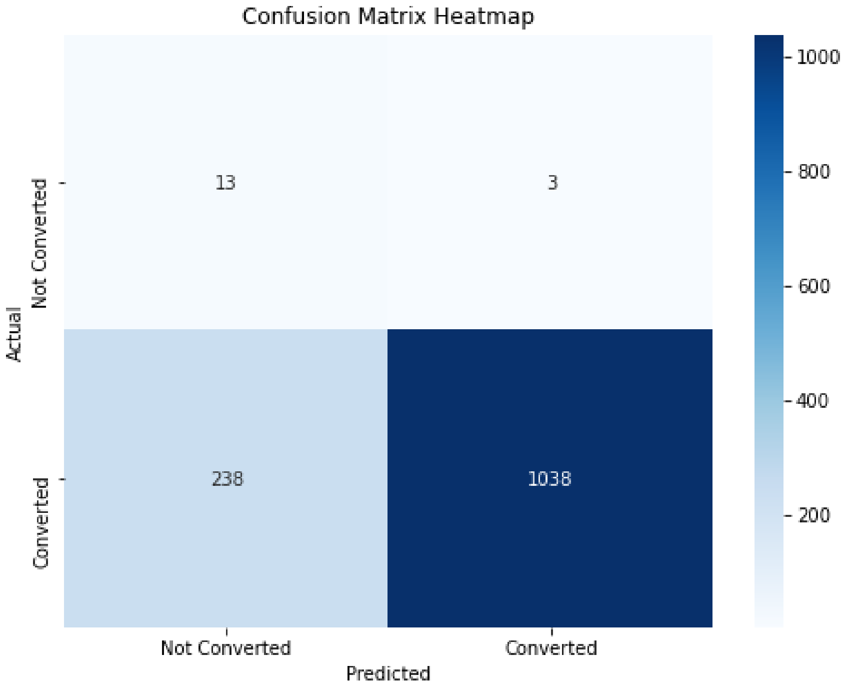
Predictive Analytics: Through logistic regression, the model quantifies the impact of various factors on the likelihood of lead conversion, offering interpretable insights that guide strategic decisions.

Machine Learning: Advanced algorithms, including cluster analysis, enable the model to uncover latent patterns within the data, segmenting leads into distinct groups based on shared characteristics. This segmentation informs targeted marketing and sales strategies, optimizing resource allocation.

Dynamic Adaptation: The model is designed to continuously learn from new data, allowing it to adapt over time and improve its predictive accuracy. This ensures that the model remains relevant and effective as market conditions and company data evolve.

Class 0 (Not Converted): Represents the leads that did not convert into sales.

Class 1 (Converted): Represents the leads that successfully converted into sales.



The adjusted confusion matrix shows a good number of TP (1038) indicating successful prediction of conversions.

The TN (13) is relatively low, which suggests that there are few cases of correctly identifying non-conversions. Given the small number of actual non-conversions (16), this isn't necessarily a problem but indicates the rarity of non-conversion cases or possible class imbalance.

FPs (3) are very low, indicating the model seldom incorrectly predicts a conversion.

FNs (238) suggests there are a considerable number of conversions that the model fails to identify. This could be an area to improve, as each missed conversion represents a potential revenue opportunity that wasn't anticipated.

A high number of TP and TN values relative to FP and FN indicates a good model performance.

A high FP rate might indicate that the model is too optimistic, seeing conversions where there are none. This might waste resources by overinvesting in leads unlikely to convert.

A high FN rate might indicate the model is too pessimistic, potentially missing out on genuine opportunities for conversion. This could mean missed revenue or lost opportunities for growth.

The high recall for the "Converted" class (0.81) indicates that the model is quite good at capturing most actual conversions, even though it's somewhat prone to missing some (indicated by the FN count). The precision for the "Not Converted" class is low (0.05), partly because the actual number of non-conversions is very small, making any error more impactful on precision and recall metrics.

Model Overview: The predictive model focuses on identifying which sales leads are most likely to convert into actual sales. The model uses historical data from our CRM and warranty systems to learn patterns that indicate a successful conversion.

High-Level Performance: Overall, the model is highly effective, with an accuracy rate of approximately 81%. This means that in 81% of cases, the model accurately predicts whether a lead will convert or not.

Strengths of the Model: One of the model's key strengths is its ability to correctly identify a large majority of the leads that will convert. Specifically, it correctly identified conversions 81% of the time, which suggests it is a valuable tool for focusing the sales efforts on the most promising leads.

Areas for Improvement: However, the model tends to miss some opportunities, as indicated by a false negative rate. This means that there are leads that could have been converted but were not identified by the model as high potential. Looking into ways to reduce this rate, which could involve further refining our data inputs or adjusting the model's parameters.

Implications for Business Strategy: The model's current performance suggests that it can significantly enhance the sales strategy by prioritizing leads with the highest likelihood of conversion. This could lead to more efficient use of sales resources and potentially higher conversion rates overall. The insights from the model also provide opportunities to revisit and optimize the lead engagement strategies, especially for those leads that the model predicts as non-converting but might still hold potential.

Next Steps: Plan to integrate the model's predictions into the sales processes, enabling more targeted follow-ups. Will continue refining the model by incorporating more detailed data and feedback from the sales team to further improve its accuracy.

Call to Action: To fully leverage the model's capabilities, I recommend a collaborative approach where sales and data teams work closely to continually refine our sales targeting strategies based on model predictions and real-world outcomes.