Contents

1 Business Context

As a key player in the Brazilian food delivery sector, in 2021, **iFood** enjoyed 86% of market share nationally, with order volume 16 times higher than its nearest competitor, UberEats (Reinaldo, 2020). As UberEats withdrew from Brazil in 2022 and Didi Food does not operate in the country, this giant is close to being a virtual monopoly (Marília, 2023).

2 Problem Formulation

With a focus on coupons (Marília, 2023), it is critical for **iFood** to address and spend marketing budget on the correct target customers which yield the most returns for the company (profit, market share, etc.). However, considering the hugely diverse product categories of **iFood**, it might be overly overwhelming and time-consuming to take into account all available product categories and their influencing factor at once. Hence, this study only focuses on a smaller with the primary goal is to generate a model being able to accurately predict the money spent on Wine products in the last two years (variable *MntWines* in the dataset) for a given customer.

3 Data Processing

'iFood_marketing.csv', which consists of multiple variables of **iFood**'s customers, is first loaded into Jupyter Notebook. Some data cleaning and processing steps are taken to deal with:

3.1 Null and duplicated values

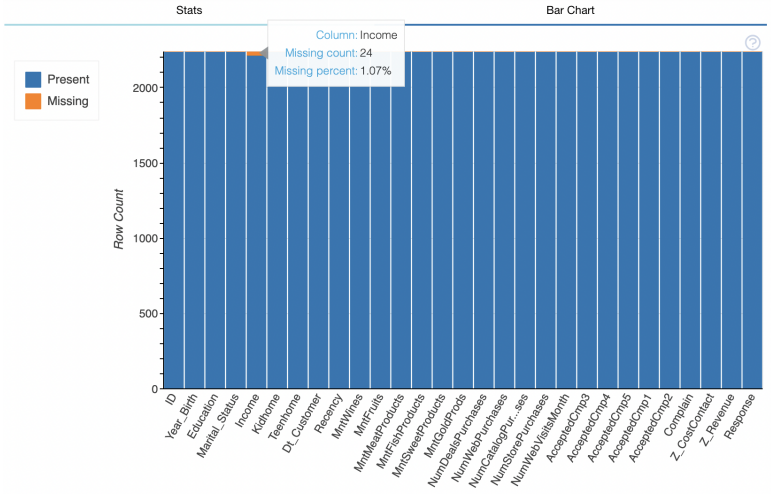

Figure 1: Null values

As can be seen, there are only 24 null values across the data set which all belong to the *Income* variables. Considering that *Income* is potentially a critical factor influencing customers' spending in general and their spending on wines specifically, those 24 rows with missing data are removed to prevent potential bias/error when being used for model building.
There is no duplicated rows thus no step is needed here.

3.2 Variables with no predicting value

Checking for unique values of each variables, it appears that each of the two columns *Z_CostContact* and *Z_Revenue* only has one value for all rows, which are 3 and 11 respectively. In other words, they yield no insights about predicting the spending of customers hence should be excluded out of the using data set.
Besides, the *ID* column only plays as a unique identified key for each customer which does not convey any insightful information about that customer. Thus, it is not deleted yet will not be included as a predictor for later analysis.

It is also noticed that under the *Marital_Status* variable, there are 4 customers with vague values ("YOLO" and "Absurd") which are hard to understand thus cannot be imputed based on available valid data. Considering the insignificant amount of those categories compared to the total data and their low interpretability, 4 rows containing those values are eliminated.

3.3 Date/Time Variables

Date/Time variables are useful for descriptive analysis yet result in many difficulties in fitting and selecting model in predictive analysis. Therefore, the *Year_Birth* column is used to created a new column *Age* - a numerical variable that directly measure a characteristic of customers which can be helpful in predicting the their spending habits. Similarly, date customer and date of birth into 2 news numerical variables. However, months instead of years. Two categorical columns are then eliminated not to cause confusion in later steps.

3.4 Categorical variables

Two columns with categorical data *Marital_Status* and *Education* are then converted into dummies using label encoding. As each variable both has more than two groups of categories, using label encoding enable creating less variables compared to binary dummies thus can avoid the 'Curse of dimensionality'. Moreover, it is obvious that *Education* is ordinal, which suits weel with the label encoding method.

3.5 Data splitting

The data set is then split into 2 separated sets: training (70%) and testing (30%). All the below analysis and model-variable selection processes are done on the training set.

4 Exploratory Data Analysis

4.1 Response: MntWines

Taking into account 1548 customers, they averagely spend $ 297.8624 on Wines products, with a standard deviation of $ 333.3153. In other words, amount of money spent on Wines products among different customers are highly dispersed. This can also be seen from the box plot below in which wide range of distribution and substantial amount of large outliners can be noticed.

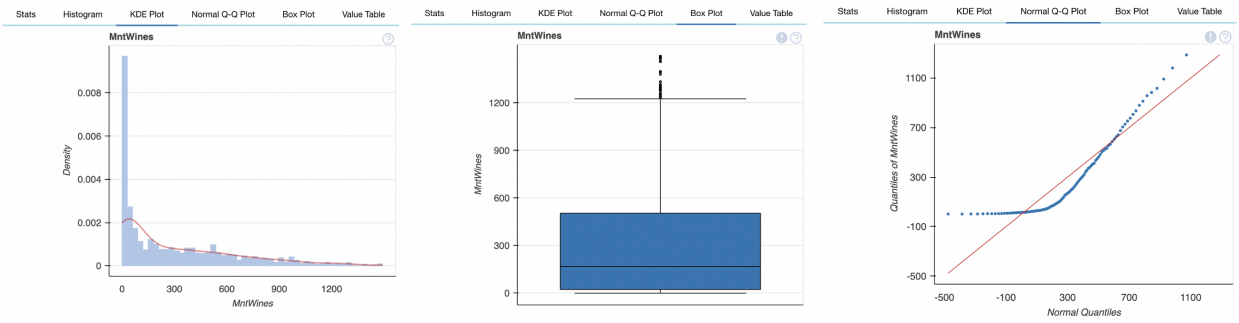

Figure 2: "MntWines" variable distribution and QQ plot

Other patterns what should also be noted include the heavy right-skewness and potential non-linearity in the distribution of the *MntWines* variable.

4.2 Response vs Other variables

Based on the heat map and correlation matrix, some variables that appear to have high correlation (≥ 0.5) with the response *MntWines* are: *Income*, *MntMeatProducts*, *NumWebPurchases*,

NumCatalogPurchases, and *NumStorePurchases*.Their relationship with *MntWines* are then further investigated using the pair plot below. As can be seen from the scatter plots, there are certain positive correlation between *MntWines* and those "interested variables" yet no clear linear patterns can be observed at this stage.

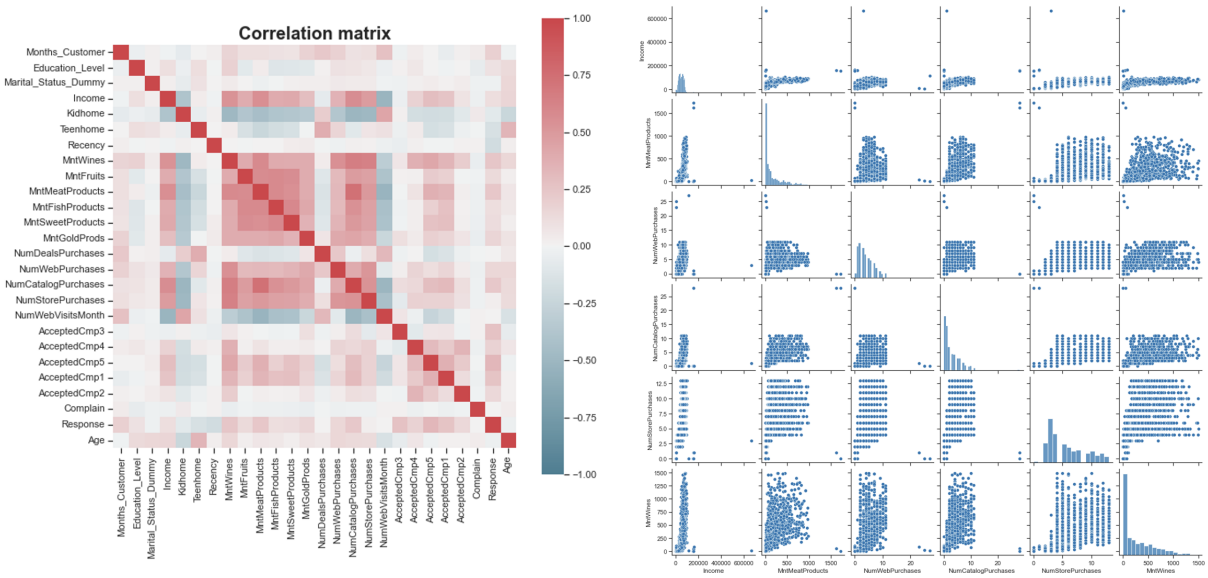

Figure 3: Correlation Heatmap and Pair Plot between variables

It is also critical to consider potential multicollinearity within the data set. When fitting regression model, if independent variables are highly correlated, the model's interpretation and accuracy can be negatively affected which can lead to overfitting. As there is no correlation coefficient that is close to 1 or -1, multicollinearity might not be a significant problem here (Appendix 1).

The distributions of *MntWines* across different groups of 2 categorical variables are also visualised which show that there are slightly difference in the average amount of money spent on Wines products among customers with different education level and marital status.

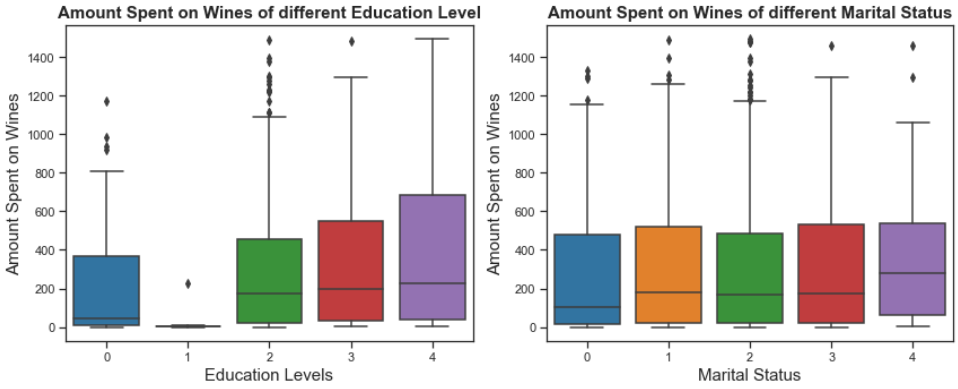

Figure 4: Distribution of "MntWines" across different categorical groups

5 Methodology

With the ultimate goal is to find a model that is accurate in predicting the money spent on Wine products in the last two years (variable *MntWine*) for a given customer, certain regression models are taken into consideration including: linear, kNN, and LASSO. To evaluate the performance of each model, Mean Absolute Error (MAE) and cross validation are used.

5.1 Linear Regression

Linear regression assumes that there is approximately a linear relationship between X and Y: $Y \approx \beta_0 + \beta_1 X$, which the training data can be used to produce estimates coefficients/parameters and predict future response values: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ (James et al., 2013). However, this model family does make certain assumptions on the data (Casson & Farmer, 2014) including:

- Linearity: The relationship between the dependent variable and the independent variables is assumed to be linear;
- Exogeneity: The residuals should be normally distributed: $E(\varepsilon|X) = 0$;
- Independence: The errors (residuals), which are the differences between the observed values and the predicted values, should be independent of each other;
- All 4th moments exist;
- No perfect collinearity: There should be no perfect linear relationship among the independent variables;
- Constant error variance: The variance of the errors should be constant across all levels of the independent variables.

For linear regression, it is crucial to efficiently select the appropriate subset of predictors. Thus, **backward selection** is used to select the most optimal subset of variables. Specifically, all potential predictors are first included in the model then gradually eliminate predictors that do not contribute to predicting the response using MAE as a measurement. In other words, if eliminating a certain predictor results in lower MAE, then the algorithm will do so. However, adding to many predictors, while lowering the error, can cause potential overfitting problem. Hence, another approach that worth being consider is the **Principle of Parsimony**, which only includes predictors with high correlation coefficients with the response (≥ 0.5)

5.2 K-Nearest Neighbours Regression

Considering the potential non-linear relationship between the response and other variables, a simple non-parametric approach such as k-Nearest Neighbours (KNN) is also a potential candidate. To select the optimal subset, a one-predictor KNN is first run with each of the variables. Then the predictor of the model with lowest MAE is chosen and two-predictor KNN is fitted with that chosen predictor and each of potential remaining predictors. For each subset of predictors, the number of nearest neighbours k is **fine-tuned** and average MAE is calculated using **cross validation**. New predictors are continuously added into the model until MAE stops decreasing or starts to increase.

In term of distance metric, as Euclidean distance only makes sense if the predictors are in the same scale, which is not the case here, **Mahalanobis distance** is chosen instead:

$$d(x_i, x_l) = \sqrt{\sum_{j=1}^{p} (x_{ij} - x_{lj})^2}$$

5.3 LASSO Regression

To deal with potential overfitting due to large amount of predictors, some regularisation methods are considered to add an additional penalty term to the formulation. In particular, as multicollinearily appears to be not a potential problem within this specific data set, Ridge regression which particularly deals with multicollinearity is not a suitable choice. Meanwhile, considering such a high dimensional data set (26 potential predictors), **Lasso regression** is chosen instead:

$$\beta_\lambda^{lasso} = \arg\min_\beta \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

As **Lasso regression** includes a penalty term that depends on the absolute values of the regression coefficients, some variables will be penalized more heavily than others simply because of their different scale. Thus, all data is standardised before being fitted to train the model. The hyperparameter (λ) is also fine-tuned using cross validation.

6 Analysis and Discussion

6.1 Linear Regression

After the **backward_selection** algorithm evaluates all potential predictors, it ends up with a model consisting of 14 predictors and the MAE equal 113.1543 (Appendix 2). However, it is noted that the p-value of *MntFishProducts* and *NumDealsPurchases* variables are greater than 0.05 (Appendix 3) which indicates that they are statistically insignificant in explaining the variation in the dependent variable *MntWines*. Thus, these 2 predictors are eliminated from the model. Another simpler model based on Principle of Parsimony, which consists of only 5 variables with high correlation coefficients with the response, is also considered:

Linear Regression Model	MAE	AIC	BIC
14 predictors	113.1543	2.061×10^4	2.069×10^4
12 predictors	113.9837	2.061×10^4	2.068×10^4
5 predictors	127.1037	2.098×10^4	2.102×10^4

As can be seen, reducing 2 predictors only slightly increase the MAE yet also reduce the BIC as well as the complexity of the model. The model with only 5 predictors, meanwhile, appears to have higher error compared to the other two. Taking this trade-off into consideration, the linear regression model with 12 predictors is chosen.

6.2 kNN Regression

As previously mention, new predictors is continuous added to the model until there is no improvements on the MAE. Their MAE and the corresponding fine-tuned *k* using cross-validation are presented in the table below:

Predictors	k-NN	MAE
"NumCatalogPurchases"	50	143.6811
as above + "NumWebPurchases"	40	125.1067
as above + "AcceptedCmp5"	45	116.0584
as above + "Income"	24	109.7661
as above + "NumWebVisitMonth"	18	103.4779
as above + "NumStorePurchases"	6	96.3127
as above + "MntMeatProducts"	6	88.7313
as above + "MntFishProducts"	4	86.0577
as above + "AcceptedCmp2"	**4**	**85.9306**
as above + "Complain"	4	86.4561
as above + "AcceptedCmp1"	4	88.6831
5 interested predictors	4	106.5377

As can be seen the MAE keeps decreasing as new predictors is added until the $10th$ predictors *Complain* is included. The model with 5 interested variables, meanwhile, yields relatively high MAE, which is understandable as KNN model is not supposed to capture the linear relationship between variables. Therefore, under KNN model family, the 4-NN model with 9 predictors: *NumCatalogPurchases*, *NumWebPurchases*, *AcceptedCmp5*, *Income*, *NumWebVisitsMonth*, *NumStorePurchases*, *MntMeatProducts*, *MntFishProducts*, and *AcceptedCmp2* is chosen (Appendix 4).

6.3 LASSO Regression

Using cross valiation, the LASSO regression chooses 16 predictors and results in a MAE of 115.9479 (Appendix 5). Those predictors that are eliminated from the model include: *Recency*, *MntFruits*, *MntFishProducts*, *MntSweetProducts*, *NumDealsPurchases*, *Complain*, *Response*, *Age*, and *Marital_Status_Dummy*.

6.4 Final Model

By comparing the MAE on the training set between three model, KNN model has the lowest error thus is chosen as the final model to predict the money spent on Wine products in the last two years.

Models	Numbers of predictors	MAE
Linear Regression	12	113.9837
KNN Regression	9	85.9306
LASSO Regression	16	115.9479

Using it on the testing data set gives a MAE of 94.6118, which is higher than the MAE on the training test yet still significantly perform better than other models. Hence, it can be concluded that the 4-NN model with 9 predictors can be considered as a relatively good predicting model for the *MntWines* variable within the scope of this study.

7 Limitation

According to **No free lunch** theorem, there is no one-size-fit-all optimal model (Ho & Pepyne, 2002) and the decision of model/variable selection should take into consideration of specific data set and ultimate goal. Hence, it is important discussing limitations of the models being discussed above before any conclusion or inference can be made:

Linear Regression: As previously mentioned, there is no clear sign on linear relationship between the response and other variables which violates the first assumption about linearity and can result in potential bias in the linear regression model. The residual plot below also suggests that the sixth assumption about constant error variance is also not satisfied. Thus, linear regression appears to be not an efficient model to predict *MntWines* variable given this data set.

Figure 5: Residual plot of linear regression model with 12 predictors

KNN Regression: Despite being the chosen model with best predicting ability, KNN regression is a memory- and time-intensive method, as it requires the entire training sample being kept in the memory for computing predictions. Thus, generating predictions using KNN regression is computationally costly. This method is also subject to a curse of dimensionality since it breaks down with high-dimensional inputs, which make it impossible to use if having larger data set with more variables.

LASSO Regression: As an extension version of linear regression, LASSO also depends on many assumptions about the data which are already discussed above. Besides, while LASSO can address high-dimensional data set, its prediction error is considered to be not as good as Ridge regression's.

In general, within the limited scope and specific requirements of this report, analysis and model/variable selection techniques are limited to linear, KNN and LASSO regression only. Therefore, the chosen model still hold relatively large prediction error (MAE = 85.9306 $) compared to the mean value of the response (297.8624 $). However, in future study in which there is no running time restrictions and, other model families (random forest tree, elastic net regression, etc.), other variable selection techniques (stepwise selection, grid search, etc.) and even other measurement matrices (root mean squared error, etc.) can also be considered and compared to reach the most optimal predictive model.

8 Conclusion

Taking into account three model families, in spite of being memory-intensive and time-expensive, KNN regression appears as the one with best predicting ability (training MAE = 85.9306 and testing MAE = 94.6118). Among all variables, *Income*, *MntMeatProducts*, *NumWebPurchases*, *NumCatalogPurchases*, and *NumStorePurchases* are the ones with highest correlation coefficients with the response variable *MntWines*.

9 Short Exercises

9.1 Best single-predictor model

Three model families above and each potential predictors are considered to find the best predictive model that uses a single predictor. The result is summarised in the table below:

As can be seen, KNN model still has the best performance with $k = 50$ and MAE = 143.8611. Using it on the testing data set results a MAE of 160.9964 which is much higher than the MAE of the KNN model with 9 predictors above.

Model	Predictor	MAE
Linear Regression	"NumCatalogPurchases"	173.5156
KNN Regression	"NumCatalogPurchases"	143.8611
LASSO Regression	"NumCatalogPurchases"	174.2739

9.2 Asymmetric errors in the prediction

Asymmetric errors in the prediction are not symmetric infers that overpredictions are substantially more risky than underpredictions. In this case, a new loss function is needed to capture this patterns, placing higher weight/penalty on overpredictions than underpredictions. Specifically, a new loss function based on the Linlin loss function suggested by Christoffersen and Diebold (1997) is used:

$$L(y,\hat{y}) = \begin{cases} \alpha|y-\hat{y}| & \text{If } (y-\hat{y}) < 0, \\ \beta|y-\hat{y}| & \text{If } (y-\hat{y}) \geq 0 \end{cases}$$

for the sake of interpretability and implementation of this report, α and β are estimated as:

$$\begin{cases} \alpha = \frac{\text{overpredicted points}}{\text{underpredicted points}}, \quad \beta = 1 & \text{If overpredicted points } \geq \text{underpredicted points}, \\ \alpha = 1, \quad \beta = \frac{\text{overpredicted points}}{\text{underpredicted points}} & \text{If overpredicted points } < \text{underpredicted points} \end{cases}$$

The reasoning behind is that when there is overprediction (number of overpredicted points is larger than number of underpredicted points), higher weight or penalty should be placed at those positive errors and vice versa. Considering the scenario where overpredictions is currently the concern, α will be positive and $\beta = 1$ as we want to tackle and minimize the overprediction phenominon.

Considering the three models in part 6.4, linear and LASSO regression tends to overpredict while KNN tends to underpredict the *MntWines* variables (Appendix 6). Thus, α is estimated as the average those of the two "overprediction" models, specifically $\alpha \approx 1.5$ and $\beta = 1$. However, it should be noted that although the high level idea of this approach aims to address the overpredictions, $\alpha \approx 1.5$ here is merely a estimate for this specific, imagined scenario thus might not yield desirable result and should be considered carefully if being used in other context.

Instead of the mean absolute error, a new metric based on this customised loss function (so-called "mean asymmetric loss error" or "weighted MAE") can be used to tackle the overprediction concern. Using the new metric, backward selection returns the same subset of predictors for linear model. Two variables with p-value higher than 0.05 are also eliminated with the same logic above. The KNN model, similarly, choose exactly the same set of predictors under the new metric, however resulting in a lower MAE value compared to the previous optimal model:

Predictors	k-NN	Weighted MAE
"NumCatalogPurchases"	50	180.6291
as above + "NumStorePurchases"	40	153.7468
as above + "NumWebVisitsMonth"	39	146.1023
as above + "MntMeatProducts"	10	133.2509
as above + "NumWebPurchases"	4	121.1657
as above + "AcceptedCmp5"	6	113.2603
as above + "MntFishProducts"	4	108.2480
as above + "Income"	4	105.5163
as above + "AcceptedCmp2"	**4**	**105.3915**
as above + "Complain"	4	105.5913

The three representative model from three model families under the new metric can be summarised as below:

Models	Numbers of predictors	MAE	Weighted MAE
Linear Regression	12	113.9837	142.4797
KNN Regression	9	66.5828	105.3915
LASSO Regression	16	115.9474	144.9349

As can be seen, evaluating by the new metric (the weighted MAE), KNN regression remains as the best predictive model. This new KNN regression model, however, even has lower MAE compared to the chosen one in part 6.4. Using this new model for the testing data set yield mean asymmetric loss error = 116.6346 and MAE = 94.6118, which is the same with the KNN model chosen under the old metric MAE.

9.3 Discussion about Fairness

Looking at the following histograms, the data for *Income* and *Age* are either heavily or relatively right-skewed with flattened upper tail. In other words, the data set only contains small amount of representatives from group of customers with high income and old age, which might potentially yield some systematic deviations (Shahbazi, 2023), or biases on the predictions when being used to train the predictive model.

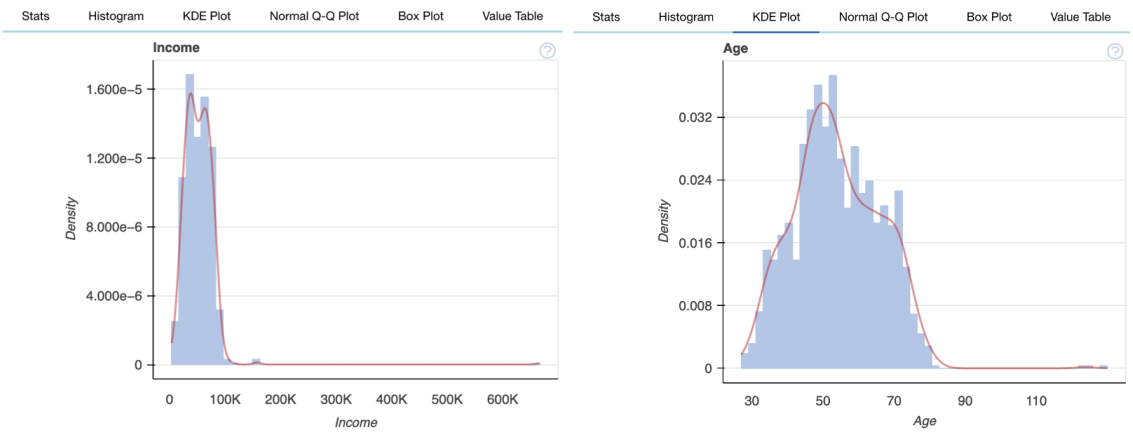

Figure 6: Distribution of customers' income and age

As the final chosen KNN model only includes a variable **Income** that can be considered 'sensitive', the existence of systematic 'bias' seems not to be a significant concern there. The linear regression model, meanwhile, contains both *Income* and *Age* which pose potential systematic 'bias'. This part, therefore, focuses on investigating whether the chosen linear regression model could be more erroneous for some groups of the population rather than others.

Specifically, customers are splitted into 3 *Age* groups: Young ($0-30$ years old), Middle ($31-64$ years old), and Old (at least 65 years old). They are also categorised based on *Income*: Low ($0-11,000$ \$), Middle (from $11,000-81,900$ \$), and High (from $81,900$ \$) (Statista, 2021). The 'protected' group is then determined as old customers with high income, which only accounts for 29 out of 1548 records of the training set:

	Young age	**Middle age**	**Old age**	**Total**
Low income	0	24	3	27
Middle income	5	1064	350	1419
High income	1	72	29	102
Total	6	1160	382	1548

The prediction performance of the linear model on protected groups is then compared that of the other groups, which is summarised in the table below:

Groups	**MAE**
Protected group	228.0414
Non-protected groups	95.5614

As can be seen, the model's predictions are consistently more inaccurate or biased for the protected groups (old customers with high income), which indicates a fairness problem.

A feasible solution that can be considered is using specific-group models, or using different personalised models for each groups. Specifically, with 3 categories for each of 2 aforementioned variables, there are 9 combinations of sub-group that can be formed, such as Old-High Income, Old-Middle Income, Young-Low Income, etc. For each of those groups, a separate linear regression can be trained using the backward selection method in part 5.1. However, it is critical to note that given the small sample size of some subgroups (Old age - Low income: 3 customers, Young age - High income: 1 customer, etc.), there might be some difficulties and potential problems (poor generalisation, high variance, etc.) in training and evaluating the corresponding models, or even impossible to do so (Young age - Low income: 0 customer). Thus, some other strategies such as collecting more data from under-represented groups, defining a new fairness metric, etc. can also be considered for future study.

10 Reference List

Casson, R. J., & Farmer, L. D. (2014). Understanding and checking the assumptions of linear regression: a primer for medical researchers. Clinical & experimental ophthalmology, 42(6), 590-596.

Christoffersen, P. F., & Diebold, F. X. (1997). Optimal prediction under asymmetric loss. Econometric theory, 13(6), 808-817.

Fioravanti, R. (2020). iFood delivers great results in Brazil: Going beyond connecting restaurants with customers. HBS Digital Initiative. Digital, Data, and Design at Harvard.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

Ho, Y. C., & Pepyne, D. L. (2002). Simple explanation of the no-free-lunch theorem and its implications. Journal of optimization theory and applications, 115, 549-570.

Marasciulo, M. (2023). A coupon-crazy Brazilian app figured out how to beat Uber Eats. Rest Of World. Rest Of World.
Shahbazi, N., Lin, Y., Asudeh, A., & Jagadish, H. V. (2023). Representation Bias in Data: A Survey on Identification and Resolution Techniques. ACM Computing Surveys.

Statista (2021). Average income in Brazil in 2021, by income percentile.

11 Appendix

Appendix 1: Multicollinearity

```
In [19]: corr_matrix[abs(corr_matrix) > 0.8].isnull().sum()
         # https://www.sfu.ca/~dsignori/buec333/lecture%2016.pdf

Out[19]: ID                    26
         Income                26
         Kidhome               26
         Teenhome              26
         Recency               26
         MntWines              26
         MntFruits             26
         MntMeatProducts       26
         MntFishProducts       26
         MntSweetProducts      26
         MntGoldProds          26
         NumDealsPurchases     26
         NumWebPurchases       26
         NumCatalogPurchases   26
         NumStorePurchases     26
         NumWebVisitsMonth     26
         AcceptedCmp3          26
         AcceptedCmp4          26
         AcceptedCmp5          26
         AcceptedCmp1          26
         AcceptedCmp2          26
         Complain              26
         Response              26
         Age                   26
         Months_Customer       26
         Education_Level       26
         Marital_Status_Dummy  26
         dtype: int64
```

Figure 7: Multicollinearity

Appendix 2: Backward selection

```
In [24]: model_fs = backward_selected(train[variables], 'MntWines', nominated=[])

         MAE if all variables included: 115.949923
         deleting Education_Level decreases MAE from 115.949923 to 114.500323
         deleting AcceptedCmp3 decreases MAE from 114.500323 to 113.948526
         deleting Recency decreases MAE from 113.948526 to 113.794792
         deleting Months_Customer decreases MAE from 113.794792 to 113.706225
         deleting Kidhome decreases MAE from 113.706225 to 113.569517
         deleting Teenhome decreases MAE from 113.569517 to 113.323161
         deleting Response decreases MAE from 113.323161 to 113.208387
         deleting MntGoldProds decreases MAE from 113.208387 to 113.186894
         deleting Complain decreases MAE from 113.186894 to 113.166330
         deleting MntFruits decreases MAE from 113.166330 to 113.159719
         deleting Marital_Status_Dummy decreases MAE from 113.159719 to 113.154271
         final model is MntWines ~ Age + AcceptedCmp5 + Income + AcceptedCmp2 + NumCatalogPurchases + AcceptedCmp1 + NumWebVis
         itsMonth + NumDealsPurchases + AcceptedCmp4 + MntFishProducts + MntSweetProducts + MntMeatProducts + NumWebPurchases
         + NumStorePurchases + 1, with MAE of 113.154271
```

Figure 8: Backward selection

Appendix 3: OLS result of MLR with 14 predictors

OLS Regression Results

Dep. Variable:	MntWines	R-squared:	0.687
Model:	OLS	Adj. R-squared:	0.684
Method:	Least Squares	F-statistic:	240.5
Date:	Wed, 27 Sep 2023	Prob (F-statistic):	0.00
Time:	00:54:43	Log-Likelihood:	-10289.
No. Observations:	1548	AIC:	2.061e+04
Df Residuals:	1533	BIC:	2.069e+04
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-305.7752	32.374	-9.445	0.000	-369.277	-242.273
NumWebPurchases	23.6564	2.208	10.716	0.000	19.326	27.987
NumCatalogPurchases	31.1803	2.709	11.509	0.000	25.866	36.495
AcceptedCmp4	190.5028	21.284	8.951	0.000	148.755	232.251
AcceptedCmp1	52.5166	22.324	2.352	0.019	8.727	96.306
Income	0.0012	0.000	4.772	0.000	0.001	0.002
MntFishProducts	-0.1733	0.122	-1.421	0.155	-0.413	0.066
Age	1.1484	0.411	2.797	0.005	0.343	1.954
MntSweetProducts	-0.4509	0.154	-2.927	0.003	-0.753	-0.149
NumStorePurchases	29.7899	2.091	14.249	0.000	25.689	33.891
NumWebVisitsMonth	17.3435	2.825	6.140	0.000	11.803	22.884
NumDealsPurchases	-4.6949	2.866	-1.638	0.102	-10.316	0.926
MntMeatProducts	0.2218	0.038	5.843	0.000	0.147	0.296
AcceptedCmp5	242.6753	23.876	10.164	0.000	195.843	289.508
AcceptedCmp2	122.2077	48.372	2.526	0.012	27.325	217.090

Omnibus:	220.586	Durbin-Watson:	2.024
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2635.909
Skew:	0.181	Prob(JB):	0.00
Kurtosis:	9.382	Cond. No.	5.98e+05

Figure 9: OLS result of MLR with 14 predictors

Appendix 4: KNN model with 9 predictors

Figure 10: KNN model with 9 predictors

Appendix 5: LASSO Regression result

	0
Income	19.5121
Kidhome	-17.3493
Teenhome	3.9151
Recency	0.0000
MntFruits	0.0000
MntMeatProducts	34.3502
MntFishProducts	-0.0000
MntSweetProducts	-0.0000
MntGoldProds	0.1557
NumDealsPurchases	0.0000
NumWebPurchases	59.0103
NumCatalogPurchases	72.1047
NumStorePurchases	78.6667
NumWebVisitsMonth	10.5066
AcceptedCmp3	7.9387
AcceptedCmp4	48.2235
AcceptedCmp5	52.1889
AcceptedCmp1	9.4528
AcceptedCmp2	9.5843
Complain	-0.0000
Response	0.0000
Age	0.0000
Months_Customer	14.4463
Education_Level	31.4029
Marital_Status_Dummy	0.0000

Figure 11: LASSO Regression result

Appendix 6: α of new loss functions

```
# linear regression
sum(fitted2 - y_train>0)/sum(fitted2 - y_train<0)
# overprediction
```

```
1.3962848297213621
```

```
# knn regression
sum(knn_fitted - y_train>0)/sum(knn_fitted - y_train<0)
# underprediction
```

```
0.8924205378973105
```

```
#lasso regression
sum(lasso_fitted - y_train>0)/sum(lasso_fitted - y_train<0)
# overprediction
```

```
1.6016806722689076
```

Figure 12: α of new loss functions of three chosen models