PREDICTING PROFITS OF INTERNET SPORT GAMBLERS USING POLYNOMIAL REGRESSION AND DECISION TREE REGRESSION

1. Introduction

Gambling, though carries certain social stigma and controversy, is one of the most lucrative and flourishing industry worldwide. In 2020, amid the influence of the Corona pandemic, the global casino industry recorded a revenue of $227 billion, and 30% of which came from online casinos [1]. With stunning increase in Internet penetration and mobile applications across the globe, the global online gambling market has been and will be an extremely promising playground for companies.

The objective of this project is to formulate and compare the efficiency of 2 machine learning models that closely predict the profit of an internet sport gambler based on his/her gambling trends. The project has a very specific business application which is to help online casinos effectively classifying customers based on their predicted profits, thereby developing specialized customer strategies such as targeted offers and loyalty program.

The flow of the project is structured as follows:

- **Section 1**: Project's real-life application and structure
- **Section 2**: Presenting the problem statement. Introducing the dataset, features and labels
- **Section 3**: Data pre-processing, feature justification and modeling
- **Section 4**: Result comparison and evaluation
- **Section 5**: Findings and room for improvement

2. Problem Formulation

Problem statement: given the data in the dataset, the supervised models aim to closely predict the profit of an internet sport gambler based on his/her gambling volume and tendency.

The dataset used in this project comes from the collaborative Internet gambling research project between the Division of Addictions (DOA) and bwin Interactive Entertainment (bwin), and was obtained from *transparencyproject.org* (title: Population Trends in Internet Sports Gambling) [2].

The dataset initially contains 46,339 datapoints, which represent the demographic and sport betting records of 46,339 users on bwin's website throughout 18 months of the research (01.02.2005 - 31.08.2006). All monetary properties are in Euros €. Originally, there are 21 features available for each datapoint, representing specific demographic and sport betting information of users such as gender, date of birth, date of registration, the total number of bets made and the total amount of money spent. Such various and insightful features are compatible for specialized predictions aimed by the models.

The 3 feature variables selected for the model are factors that highly reflect the gambling trend of a sport gambler including the total number of bets, the total number of days doing sport betting on bwin's website and the average amount of money spent on a bet. One of these 3 features (the average amount of money spent on a bet) was not originally available in the dataset, and operations done to produce it will be

explained thoroughly in section **3.1**. The label is the profit of a sport gambler in monetary value (regression).

3. Methods

3.1 Dataset

Due to limitations in memory capacity, I have reduced the size of the dataset to 8000 datapoints. The dataset was pre-processed by removing of irrelevant columns, renaming important columns for easier understanding, and filtering datapoints that represent inactive users (users that did not make any sport bet during the research period). 15 irrelevant features were removed from the dataset, leaving 6 important features. Specific descriptions of those 6 features are presented in the Table 1.

Table 1

Feature	Description	Data type
UserID	ID number bwin assigned to each user upon registration	Numeric
BirthYear	Birth year of the user	Numeric
TotalSpendings	The total amount of money the user spent on bets	Numeric
TotalEarnings	The total amount of money the user won from bets (note that this is not equal to profit). For example, if a user spends €10 on a bet and earns €18, his/her profit on that bet is €8; while if a user spends €10 a bet and earns €0, his/her profit on that bet is €-10.	Numeric
Bets	The total number of bets the user made	Numeric
Days	The total number of days the user did sport betting on bwin's website	Numeric

Importantly, 2 new columns were added into the dataset using pandas library [3], which are *AverageBet* (the average amount of money that a user spent on a bet, which can be calculated by the formula *TotalSpendings/Bets*) and *Profit* (the total amount of money a user profited from sport betting, which can be calculated using the formula *TotalEarnings - TotalSpendings*).

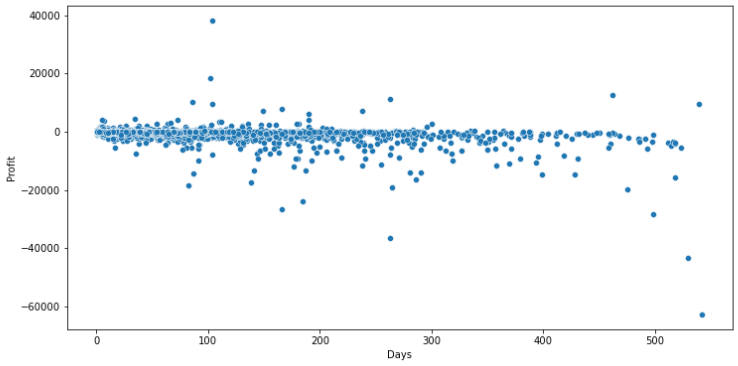
	UserID	BirthYear	TotalSpendings	TotalEarnings	Bets	Days	AverageBet	Profit
0	1324354	1963	17896.8500	16820.8500	798	233	22.427130	-1076.0000
1	1324355	1983	663.9998	580.5700	307	101	2.162866	-83.4298
2	1324356	1977	1600.6300	985.6300	242	54	6.614174	-615.0000
3	1324358	1981	336.2897	209.8574	11	5	30.571791	-126.4323
4	1324360	1978	692.6065	626.9512	394	60	1.757885	-65.6553
5	1324362	1983	283.4200	216.2400	176	57	1.610341	-67.1800
6	1324363	1983	41.5300	0.0000	3	3	13.843333	-41.5300
7	1324364	1968	562.0000	293.6500	113	46	4.973451	-268.3500
8	1324368	1983	8246.4600	8076.4000	497	89	16.592475	-170.0600
9	1324369	1985	1535.4313	1445.8534	525	109	2.924631	-89.5779

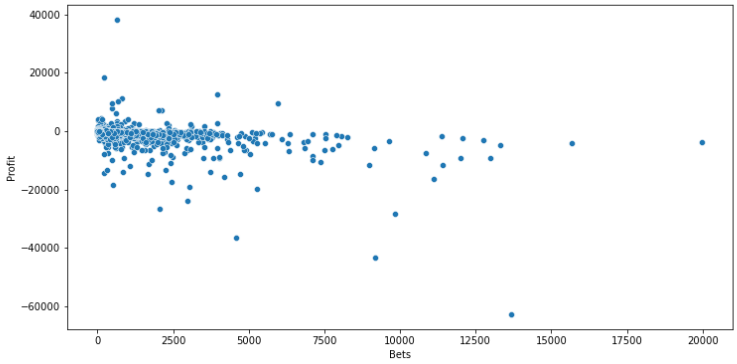
Feature Justification

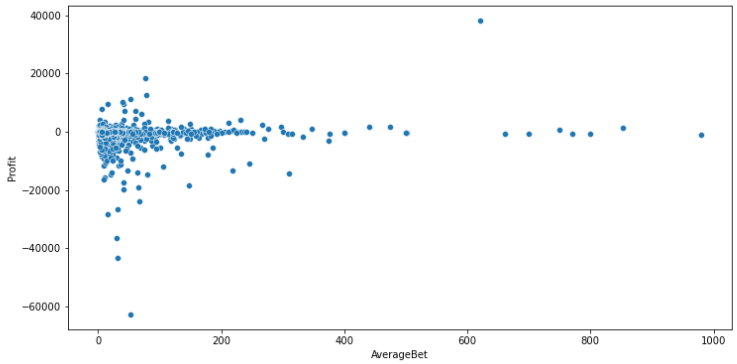
- *Days*: The number of days that a user plays on bwin's websites greatly reflects the user's gambling frequency and gambling urge. There's a high chance that a user who plays in many days has a strong and continuous desire to gamble, while a user who plays in a handful of days only gambles if a big sport event is ongoing.
- *AverageBet*: This feature reflects the gambling tendencies of a user, that is, how much money that user is willing to spend on a bet, averagely.
- *Bets*: This feature shows the user's gambling volume, and when combined with data in *AverageBet* can illustrate the user's gambling strategy (for example: playing lots of small bets or playing a handful of very big bets)

The combination of these 3 features provides a very strong reflection of a user's gambling volume and tendency, making them applicable feature variables for the specialized predictions aimed by the models.

I also visualize the data using scatterplot, and 3 features show a relatively strong correlation to the label.







When it comes to splitting the dataset into training, validation and test sets, K-fold cross-validation is especially useful in case there is a relatively small dataset where the risk of the single split is relatively high. Since the dataset used in this project contains of 8,000 data points, I figured it was not necessary to use k-fold cross validation to estimate the optimal amount of data to use to avoid under or overfitting. The dataset was split into 3: the training set, the validation set, and the test set with the ratio of 0.6:0.2:0.2, respectively, as when testing with multiple different ratios, this split yields that best results (lowest errors).

3.2 Polynomial regression

Visualization of the data with scatterplot shows curve-linear relationships between the 3 features and the label. Therefore, polynomial regression is a reasonable choice. The mathematical ground of polynomial regression of k-th degree is:

$$\hat{y} = \beta_0 + \beta_2 * x_1^1 + \beta_2 * x_1^2 + \ldots + \beta_n * x_n^k$$ [4]

In regards to loss functions, PolynomialFeatures only supports LinearRegression class, in which the customary choices of loss function for polynomial regression are mean squared error (MSE) and mean absolute error (MAE). While MSE is usually used for training and validation error roots to make large numbers more readable, MAE is a lot more robust to outliers. Since MSE squares the error *(y — y_predicted = e)*, its punish rate increases exponentially and if there exists an outlier in the data, the value of *e* will be high and e^2 will be significantly larger than $|e|$. This will make a model with MSE loss give more weights to outliers than a model with MAE loss. Since the chosen dataset for this project contains a

considerable number of extreme outliers, MAE loss function is likely to be the superior choice. The mathematical ground of mean absolute error is:

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n}. \quad [5]$$

3.3 Decision tree regression

Even though it was observed in the visualization that the 3 features have curve-linear relationships with the label, the dataset contains a considerable number of extreme outliers, which can heavily affect the accuracy of the polynomial regression model. Because of its competency in analyzing both linear and non-linear relationships and its robustness to outliers, decision tree regressor is adopted as the second method. All tree algorithms are robust to outliers as they operate by the principle of splitting the dataset into smaller groups in a multidimensional space considering the minimum sum of squared residuals and they do not create a hyperplane like linear regression, polynomial regression or support vectors.

In regards to loss function, mean absolute error (MAE) is also used for this model. The motivation for this is similar to polynomial regression's choice of loss function, as MAE is one of the customary loss functions for decision tree regression and is robust to outliers.

4. Results

4.1 Results of polynomial regression model

Choosing the optimal degree is one of the most important aspects when using polynomial regression. Correct degree selection is vital since it directly influences the assumed relationship between the selected features and the label value. Performance of the polynomial regression model on the training set and the validation set is assessed by comparing the training error and the validation error of polynomial regression with the degree ranging from 2 to 6.

Table 2. Training error and validation error of polynomial regression of degree 2 to 6

Degree	2	3	4	5	6	7
Training Error	324	264	241	288	307	340
Validation Error	360	328	388	2514	2872	19789

Comparing the training error and validation error of different degrees, the optimal degree is likely to be 3, since it yields the lowest validation error (328) and an acceptable training error (264). From degree 5 onwards, the model seems to start overfitting heavily.

4.2 Results of decision tree regression model

Choosing the optimal tree depth is one of the most important aspects when using decision tree regression, as assigning the unsuitable depth for the model might result in overfitting or underfitting. Performance of the decision tree regression model on the training set and the validation set is assessed by comparing the training error and validation error of decision tree regression with the max depth ranging from 1 to 8.

Table3. Training error and validation error of decision tree regression with max depth of 1 to 8

Degree	1	2	3	4	5	6	7	8
Training Error	344	330	292	267	245	225	199	174
Validation Error	403	396	360	344	331	338	349	376

Comparing the training error and validation error when assigning different max depths, the optimal max depth is likely to be 5, as it yields the lowest validation error (331) and an acceptable training error (245).

4.3 Comparing 2 methods

Comparisons of the training error and validation errors obtained by the 2 models show that there are no significant differences on their performances on the training and validation sets. While the performances of the 2 models on the validation set are relatively similar, the decision tree regression model performs slightly better on the training set with the training error of 245 (compared to the polynomial regression model's training error of 264), making it the slightly superior choice overall. Thus, decision tree regression with max depth of 5 is selected as the final method.

The competency of the decision tree regression model is finally assessed by letting it perform on the test set. Since the test set contains datapoints that are completely new to the model, it is a valuable metric to assess the model's competency. As stated on section **3.1**, the test set accounts for 20% of the dataset, containing 1600 datapoints. For the final chosen method, the test error is 268.

5. Conclusion

Through detailed assessment of 2 models' performances on the training and validation sets, decision tree regressor model has been demonstrated to be the slightly superior option. This might be due to the appearances of multiple extreme outliers in the dataset, which can heavily affect the accuracy of polynomial regression model. However, polynomial regression model also performs surprisingly well as it yields very close errors to decision tree regression model.

In this project, it is noticeable that we have relatively large errors and 2 reasons could account for them. Firstly, as sport betting is based on real-life sport competitions, many unexpecting events might happen (e.g., a semiprofessional football team beats the national champion), which results in large variance in players' profits. Such unpredictability is the fundamental reason why most people gamble – you never know for sure what is going to happen, and the bookmakers also do not know. Secondly, it is because that the currency it is Euros £, as we are trying to predict the profits of sport gamblers in monetary value.

Although the results seem to be promising enough, there is still room for improvement. Outliers represent anomalies that are extremely worth-noticing for online casinos and should be detected, since they represent customers that either cause the most damage or bring the most profit. Therefore, the outliers in

the chosen dataset deserve more meticulous treatment. As promising directions for future work, I consider defining boundaries for outliers and imputing the outliers using a variety of imputation values (e.g., features, median, mode, or boundary values). Another side topic would be to investigate which other features might affect a gambler's profit. Although the chosen feature variables are reasonable choices, it would be interesting to see if demographic factors (e.g., age, gender, and country of residence), for example, places a significant influence on the gambler's result.

6. Reference

[1] 10 Online Gambling statistics for 2021 and Beyond
Available at: https://playtoday.co/blog/online-gambling-statistics/

[2] Dataset: Population Trends in Internet Sports Gambling
Available at: http://thetransparencyproject.org/Availabledataset.htm

[3] Pandas library
Available at: https://pandas.pydata.org/

[4] Polynomial regression
Available at: https://en.wikipedia.org/wiki/Polynomial_regression

[5] Mean Absolute Error
Available at: https://en.wikipedia.org/wiki/Mean_absolute_error