Implementation of Statistical Tool in Legendary Pokémon Analysis and the Pricing in Stock Market

Lin Qiao[1, †] and Yangning Tan[2, *, †]

[1]Faculty of Science and Technology, Beijing Normal University-Hong Kong Baptist University United International College, Zhuhai, China

[2]School of Mathematical Sciences, Shanghai Jiaotong University, Shanghai, China

*Corresponding author: tanyangning@sjtu.edu.cn

†These authors contributed equally.

Abstract. Statistics are widely applied in wide spread of advanced science researches to give the quantitative crustily, which is now much more well-liked. In this paper, a dataset is analyzed and a pricing method is presented based on two different statistical models. In the random forest model, we first do a simple statistical data analysis to determine the association between attributes and "is legendary" using the Legendary Pokémon dataset. The above analysis results may then be used to determine which characteristics can make a Pokémon legendary. In the Black-Scholes model, we first discuss SDE before using it to provide a stock pricing method. For the former one, the statistical data analysis in the first application may determine that all the characteristics have a positive correlation with "is legendary," and lastly, random forest discovered that the characteristics that can turn a Pokémon legendary are special attack and speed. Through the second application, it can determine how an SDE's trajectory and coefficients are related. The pricing method is then obtained by applying SDE to the Black-Scholes model. We value a company's shares using this method and contrast the accuracy of this approach with more conventional approaches. According to data analysis, the first application easily and intuitively demonstrates random forest's practical application capabilities, as well as its ease and accuracy for us. These results shed light on guiding further exploration of the practice of random forests in data analysis and the application of SDE related models.

Keywords: Statistical tool; Pokémon; Random forest; SDE; Black-Scholes model.

1. Introduction

A commonly used statistical model in many different fields is called random forest. The random forest model was employed by Farnaaz and Jabbar as the Intrusion Detection System's classifier (IDS). This model outperforms other conventional classifiers in successfully categorizing assaults, with a reduced false alarm rate and a better detection rate [1]. RF is a highly common technology in the Remote Sensing Industry. It may examine how RF classifiers are used in remote sensing. The Variable Importance (VI) assessment produced by the RF classifier has been extensively employed in several contexts, which is the most crucial factor. For instance, choosing the best season to categorize target categories, narrowing the dimension of hyperspectral data, and determining the most pertinent geographic and remote sensing data from several sources [2]. Additionally, machine learning technology is being used more often in modern biology for the sophisticated and extensive examination of biological data. Recent work in Bio-informatics has witnessed a growth in the usage of RF because of its special benefits in processing tiny samples, high-dimensional feature space, and complicated data structures [3].

Another popular statistic tool is Stochastic Differential Equation (SDE). SDE is widely applied in financial systems, quantitative economics, and control systems. In models of financial system, for example, SDE and the parameters in it are studied to optimize the model. Therefore, there has been researches about the theory of SDE [4]. Among all the models, the most famous and popular one is the Black-Scholes model [5]. The Black-Scholes model can be used in market pricing with the advantage of being relatively accurate [6]. By collecting data from the past, companies can price their stocks in a reasonable way and buyers are able to make decisions in the stock market based on the data. This is a useful tool in finance for both companies and individuals.

In addition to its extensive usage and advancement in the aforementioned three disciplines, random forest is also frequently employed as a classifier in statistical data analysis. A few academics have also released pertinent studies in the previous two years. For instance, Clark attempted to stratify Pokémon in 2020 by doing a number of statistical analyses on the degrees of competitive Pokémon fights and developing a logistic regression model and a random forest model. To determine which layer Pokémon, belong to, these models will consider the fundamental statistical total, all statistical values, the Pokémon's generation, whether it is a legendary Pokémon, and if it has undergone super evolution [7]. Subsequently, in 2021, taking into account statistical data and different Pokémon kinds, Navas and Donohue looked at the hierarchical classification employed by competitive Pokémon. They employed the Weka J48 decision tree, slow IBk 1-nearest neighbor, and logistic regression algorithms for classification and assessed the methods using accuracy and precision metrics. The findings of this study provide light on the variables that have the greatest influence on the deployment of Pokémon by competitive teams and may be used to predict how Pokémon will fare in the future when other Pokémon lose their appeal or vanish from the game and other Pokémon are added [8].

Regarding to SDE and the Black-Scholes model, they originated from the need of mathematical modeling in financial field. In 2019, Dunbar recapitulated the development in SDE-related financial model, from Brownian motion to stochastic calculus and finally to the Black-Scholes model [9]. However, the Black-Scholes model itself is a demanding research topic with SDEs. In order to solve this problem, different methods are applied. In a 2020 paper, Saratha uses a method called fractional generalized homotropy analysis method (FGHAM) [10]. It is able to avoid the problems in applying multiple integration and differentiation procedures while obtaining the solution with well-established methods. The result of the study has enabled us to use the Black-Scholes model to price the stock of a company in more complex situations. To be precise, it is able to make a prediction with non-constant coefficients in the SDE. Therefore, the model can be optimized to fit the real-life situation and give an estimation of higher accuracy.

As a result, statistical techniques are crucial in all professions. This research aims to analyze how random forest and the Black-Scholes model are used. To determine what qualities define a legendary Pokémon, this study first analyzes the Pokémon dataset using a random forest model. We must first do a fundamental statistical data analysis to determine whether there is a positive or negative association between the qualities and legendary stars before applying the random forest model. Then, the Variable Importance (VI) measurement of the random forest model can intuitively show us the attributes that contribute the most to legend. Combined with the correlation obtained by statistical analysis, we can answer our research questions. To price the stock of a company, this research gives a general method using the Black-Scholes model. The first thing to talk about is SDE because SDE is a key concept in the Black-Scholes model. With both numerical and analytical methods, we explain the role of the coefficients in SDEs. Afterwards, a general description of how we obtain the Black-Scholes model is given with suppositions and deductions. Finally, it is tried to use the Black-Scholes model to price the stock of a company, proving its reliability and accuracy to some extent.

2. Legendary Pokémon Analysis

Video game Pokémon is highly well-liked. Several Pokémon in the game is legendary, and they are often incredibly strong. Therefore, the goal of this application is to find one or two attributes that are most likely to make Pokémon legendary. The experiment started by examining the statistical data for various Pokémon attributes and found that the legendary Pokémon is more potent overall. The significance of the variables are ranked using the random forest algorithm, with legendary serving as the classification variable, in order to more properly examine. Our dataset about Legendary Pokémon, which obtained from Kaggle website [11], includes a classification of Pokémon types along with several numerical fighting attributes, including as attack, defense, speed, etc. The binary classification variable is legendary, marked 1 if a Pokémon is a legendary and 0 otherwise, is what we will ultimately be predicting. The research anticipates conducting an analysis on the dataset's sole binary classification variable. To investigate attributes that are most likely to cause "is legendary = 1", the research must first determine if there is a positive or negative correlation between the variables in order to answer this problem. As a result, a few statistical data analyses were conducted for numerical variables.

2.1 Statistical Data Analysis

For all 801 Pokémon, the research illustrates the link between height m and weight kg, emphasizing the legendary Pokémon. As shown in the Fig. 1, the blue points on this graph represent legendary Pokémon and with a few notable exceptions, legendary Pokémon are often heavier and taller. For instance, despite their great heights, Onix, Steelix, and Wailord are not legendary characters. There must be other factors at play. As shown in the figure, the proportion of legendary Pokémon in different types and present the results using a simple bar chart. Pokémon types differ significantly in how they relate to legendary status. There is no legendary poison or fighting Pokémon, despite the fact that more than 30% of flying and psychic Pokémon are legendary. The box plots in Fig. 1 consider the influence of a Pokémon's fighter stats (attack, defense, etc.) on its status simultaneously. Legendary Pokémon outperform their common counterparts in every fighting stat, as one might anticipate. The boxplots indicate a significant difference for each of the six variables, even though we haven't explicitly tested for a difference in means. Despite this, each instance has a handful of outliers, which means that certain legendary Pokémon are abnormally weak.

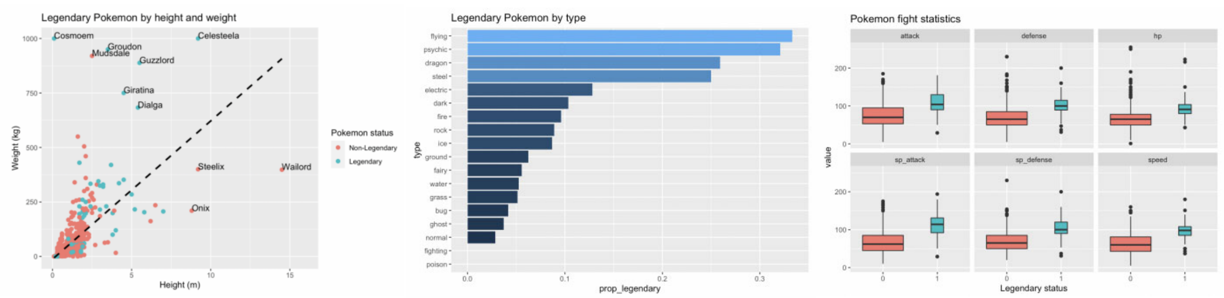

Fig. 1 Statistical Data Analyses: Scatter Plot, Bar Chart and Box Plot (from the left to the right)

2.2 Random Forest

Before fitting the model, the dataset was divided into two parts: training and testing. Then, after running the code, the results show that the OOB estimate of error rate is 6.17%, and the confusion matrix is shown in Table. 1. The confusion matrix shows the full categorization of each group. For the class of not legendary, 420 samples are categorized as not legendary, whereas only 5 samples are classed as legendary, resulting in an error of around 0.012. Regarding to the class of legendary, 21 samples are categorized as legendary, while 24 samples are classed as not legendary, resulting in an error of roughly 0.533. Then, the ROC curve for the model can be plotted to compute the Area Under the Curve (AUC) and visualize the True Positive Rate (TPR) and False Positive Rate (FPR). The closer the curve is to the top left of the plot, the higher the Area Under the Curve (AUC) and the better the model. AUC curve is shown in the Fig. 2. Through calculating, the result of the AUC is nearly 95%, which means the model is useful to determine variable significance.

Table 1. Confusion Matrix

	Confusion Matrix		
	0	1	Class Error
0	420	5	0.01176471
1	24	21	0.53333333

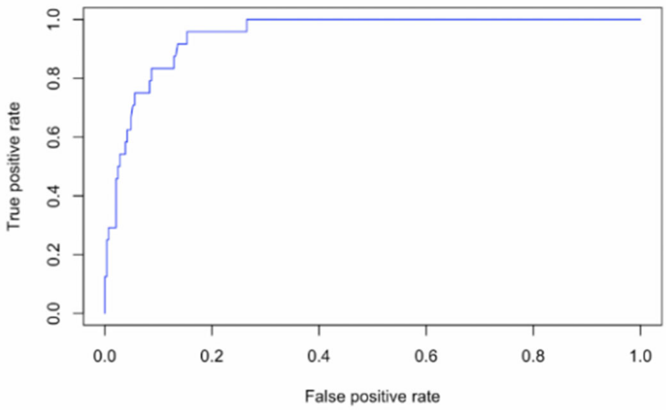

Fig. 2 ROC Curve

2.3 Variance Importance

In order to find out which attributes make Pokémon legendary, two indicators of the significance of a variable from Random Forest need to be found. MeanDecreaseAccuracy is the amount that omitting a certain variable reduces the model's accuracy. MeanDecreaseGini measures how much a variable increases the likelihood that an observation will be categorized in one of two ways (also known as "node purity"). The measurement results of Variable Importance based on these two indicators are shown in Table 2 and Fig. 3.

Table 2. Value of Variance Importance

	Mean Decrease Accuracy	Mean Decrease Gini
attack	10.900305	5.649163
defense	5.615223	4.575497
height_m	11.430842	10.471343
hp	14.894096	7.439156
sp_attack	14.977600	12.079566
sp_defense	11.519140	7.996919
speed	16.931984	10.670712
type	6.833440	8.697711
weight_kg	12.444648	9.849826
generation	3.189315	3.288933

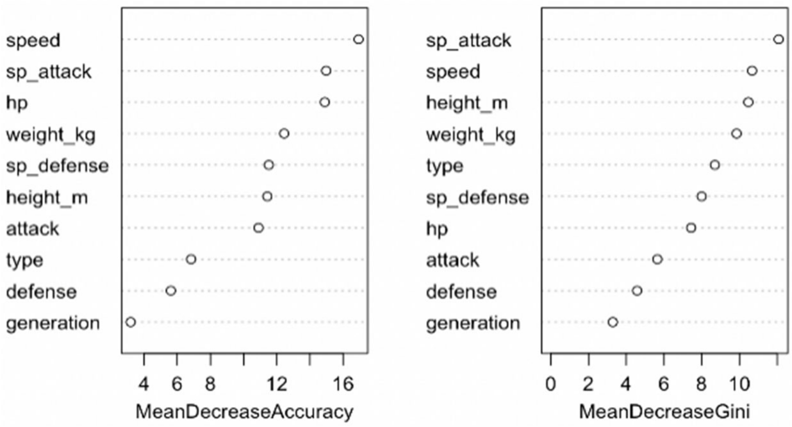

Fig. 3 Variance Importance Plot

Both the significance of the variables of these two indices increases as the value of the variable increase. Together, these two measures will allow us to answer our original research question what makes a Pokémon legendary. Therefore, special attack is the most essential component in determining whether a Pokémon is legendary, followed by speed. Combined with the results of previous statistical data analysis, we can determine that the higher the value of special attack and speed, the more likely the Pokémon is to be legendary. As a result, the legendary Pokémon are distinguished firstly by the strength of their special attack and, secondarily, by their speed, while also displaying superior combat ability across the board. However, as previously stated, there are still some legendary Pokémon with low special attack and some non-legendary Pokémon with strong special attack. In order to solve this problem, the research supposed to extend and enhance this classifier by looking at additional Pokémon attributes such as catch rate, location, age, or gender rate.

3. Implementation of Statistical Tools in the Pricing of the Stock Market

3.1 SDE

One of the problems widely existing in specific models based on SDE is what information each coefficient in the SDE provides. Therefore, the situation in which the coefficients are constant is studied to find out the relationship between the coefficients and the trajectory of an SDE.

$$\begin{cases} dX_t = \alpha(v - X_t)dt + \sigma dB_t \\ X_0 = x_0 \end{cases} \tag{1}$$

3.1.1 Coefficients and the trajectory of SDE

First, trajectories with different coefficients will be plotted to study this problem. The first question is what σ stands for. Other coefficients are fixed to see what feature of the trajectory different values of σ reveal. By plotting the trajectories (seen from Fig. 4), when σ takes a greater value, the volatility of the trajectory tends to be greater as well. Therefore, one can assume that σ is the coefficient indicating the volatility of the trajectory. Subsequently, another coefficient v will be checked. Similarly, other coefficients are fixed to examine the effect of v on the trajectory (seen from Fig. 5). The tendency of the trajectory to converge can be seen clearly. Moreover, the trajectories seem to converge at the exact values of v. Although still oscillating because of the Brownian motion, the convergency is obvious. Therefore, it is assumed that v is the convergent limit of the trajectory.

Finally, α will also be talked about. Likewise, all other coefficients are fixed to find out the effect of α on the trajectory. In the study of v, the fact that the trajectory converges has already been noticed. With different values of α, one can also see that when α takes a larger value, the trajectory converges more quickly. Thus, it can be assumed that α stands for the rate of convergence of the trajectory as depicted in Fig. 6.

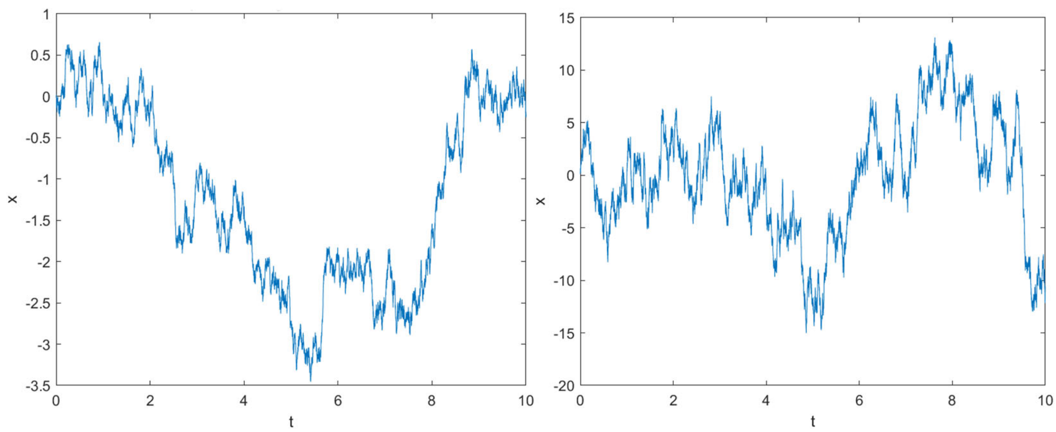

Fig. 4 The trajectory of SDE with different σs

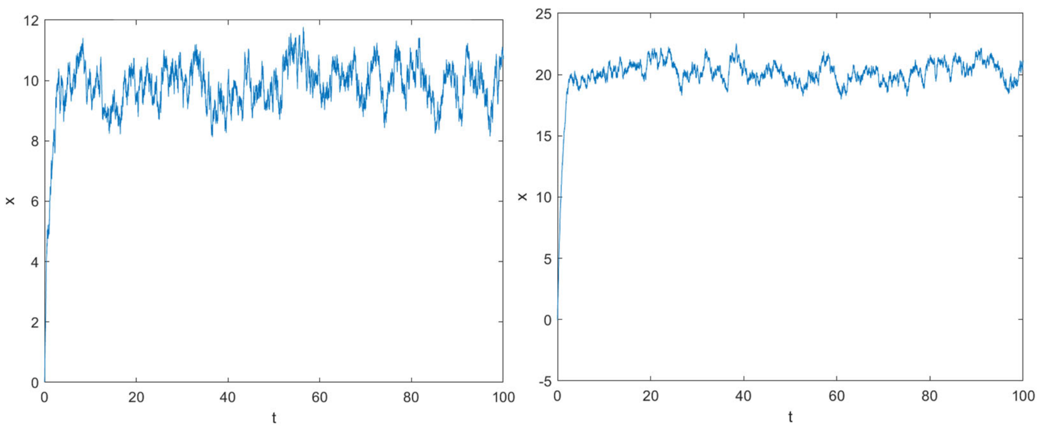

Fig. 5 The trajectory of SDE when v = 10 and v = 20

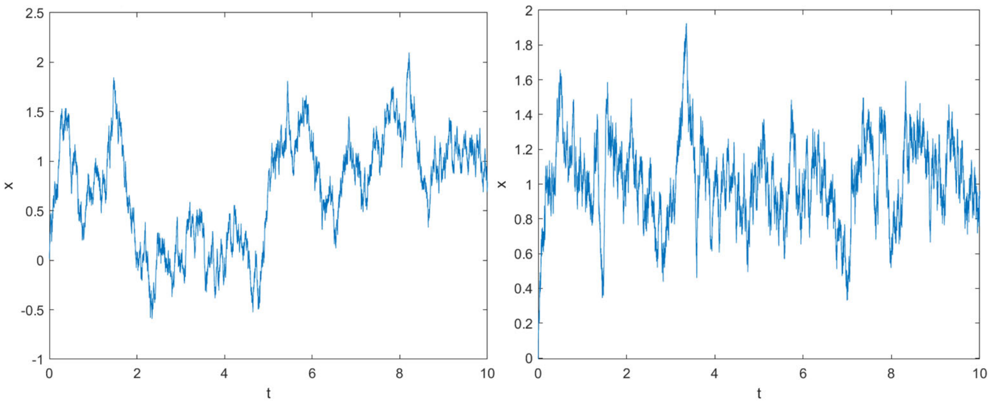

Fig. 6 The trajectory of SDE with different αs

3.1.2 Solution of SDE

Given an initial value, an SDE can be solved through calculation. The SDE is given in Eq. (2).

$$\begin{cases} dX_t = \alpha(v - X_t)dt + \sigma dB_t \\ X_0 = x_0 \end{cases} \tag{2}$$

Subsequently, it can be solved with Ito-Doeblin formula in Eq. (3).

$$d(e^{\alpha t}X_t) = e^{\alpha t} \cdot (dX_t + \alpha X_t dt) = e^{\alpha t} \cdot (\alpha v dt + \sigma dB_t) \tag{3}$$

With the initial value, the solution can be given in Eq. (4).

$$X_t = e^{-\alpha t} \cdot \left(x_0 + \alpha v \int_0^t e^{\alpha s} ds + \sigma \int_0^t e^{\alpha s} dB_s\right) = x_0 e^{-\alpha t} + v(1 - e^{-\alpha t}) + \sigma e^{-\alpha t} \int_0^t e^{\alpha s} dB_s \tag{4}$$

Now that the exact solution of the SDE is given, it is able to check the relationship between the coefficients and the trajectory of the SDE. To start with, σ is the coefficient of the Brownian motion. Therefore, the greater the value of σ is, the greater the volatility is. Meanwhile, it is noticed that when t tends to infinity, the X_t tends to v, which means that v is the limit of convergence of the trajectory. Finally, it is also clear that when α takes a greater value, X_t converges to v more quickly. Therefore, α stands for the rate of convergence.

3.2 The Black-Scholes Model

3.2.1 Introduction of the Black-Scholes Model

In the Black-Scholes model, there are two assets, the stock price and the debenture bond. Here, S_t is used to represent the stock price and B_t is used to represent the debenture bond. In addition, the Brownian motion is represented by W_t to avoid confusion. The stock is considered no dividend and therefore the share price of the stock follows the SDE with constant coefficients shown in Eq. (5).

$$dS_t = \mu S_t + \sigma S_t dW_t \tag{5}$$

Meanwhile, the profit of the debenture bond is supposed to be a continuous compounding with a constant rate of interest r. Hence, B_t follows the ODE shown in Eq. (6).

$$dB_t = rB_t dt \tag{6}$$

Combining these two assets, we have an asset portfolio Π_t, i.e., Π_t can be represented by B_t and S_t in Eq. (7).

$$\Pi_t = \gamma_t B_t + \Delta_t S_t \tag{7}$$

Moreover, the Black-Scholes model supposes that the asset portfolio is self-financing. Therefore, by taking derivative of Eq. (7), one can get Eq. (8).

$$d\Pi_t = r\gamma_t B_t dt + \Delta_t dS_t \tag{8}$$

The price of the financial derivative is denoted by V_t. Consequently, there is $V_t = \Pi_t$ according to the arbitrage-free principle. Therefore, there also is $dV_t = d\Pi_t$. Again with Ito-Doeblin formula, one obtains Eq. (9).

$$dV_t = \dot{V}_t dt + \dot{V}_x dS_t + \frac{1}{2}\ddot{V}_t d[S_t, S_t] = \left(V_t + \frac{1}{2}\sigma^2 S_t^2 \ddot{V}_{xx}\right) dt + \dot{V}_x dS_t \tag{9}$$

Using $dV_t = d\Pi_t$ and comparing the coefficient of dS_t and dt, one can get Eq. (10).

$$\begin{cases} \dot{V}_x = \Delta_t \\ \dot{V}_t + \frac{1}{2}\sigma^2 S_t^2 \ddot{V}_{xx} = r\gamma_t B_t \end{cases} \tag{10}$$

With the terminal condition $V_T = (S_T - K)^+$, the solution to the Eq. (10) is shown in Eq. (11):

$$V_t = SN(d_1) - Ke^{-r(T-t)}N(d_2) \tag{11}$$

Here,

$$d_1 = \frac{\ln\frac{S_t}{K} + \left(r + \frac{1}{2}\sigma^2\right)(T-t)}{\sigma\sqrt{T-t}}, d_2 = d_1 - \sigma\sqrt{T-t} \tag{12}$$

3.2.2 Implementation of the Black-Scholes Model

Now, an example is given to show how the Black-Scholes model is used to price the stock of a company. It is supposed that a company has a five-year bond with a value of \$37.5 million and an annual interest rate of 5.5%. In calculation, the expected rate of interest is 6% without considering the risk and 11% after adjusted with respect to the risk. Other bonds of the company have a value of \$87.5 million and an annual interest rate of 8%. The probable value of the company and the probability is given in Table 3.

Table 3. The value of the company in five years

The value of the company	150	250	400	600	900
probability	0.1	0.2	0.3	0.3	0.1

To price a stock with the Black-Scholes model, the current market value of the company S and the volatility σ are needed. In this case, the expectation of the future value will be used to estimate the current value. The probability of different S_i is shown in the table 4. With simple calculation, one can get the expectation of future value of the company, which is \$326.5 million. Since the interest rate adjusted according to the risk is 11%, the current price of the company should be $K = \frac{326.5}{(1+11\%)^5} = 193.76$. Denote $u_i = \ln\frac{S_i}{K}$, the value of σ can be estimated with the standard deviation shown in table 5.

Table 4. S_i and the probability

S_i	21.5	121.5	271.5	471.5	771.5
probability	0.1	0.2	0.3	0.3	0.1

Table 5. Standard deviation

p_i	S_i	u_i	s_i^2
0.1	21.5	-2.2	0.572
0.2	121.5	-0.47	0.087
0.3	271.5	-0.337	0.006
0.3	471.5	0.889	0.145
0.1	771.5	1.328	0.142

From the Table. 5, there is $s = \sqrt{\sum s_i^2} = 0.976$, $\sigma = s/\sqrt{t} = 43.6\%$. Now, one can use the Black-Scholes model to price the stock. d_1 and d_2 are given in Eq. (13).

$$d_1 = \frac{\ln\frac{S}{K} + \left(r + \frac{1}{2}\sigma^2\right)T}{\sigma\sqrt{T}} = 2.205, d_2 = d_1 - \sigma\sqrt{T} = 1.230 \tag{13}$$

Thus, $N(d_1)$=0.9863 and $N(d_2) = 0.8893$. Finally, the price of the stock is $V_t = SN(d_1) - Ke^{-rT}N(d_2) = 158.82$. Therefore, if the company issues ten million shares of the stock, the price of each share should be \$15.88. With a traditional method, the binomial tree model, one can also price the stock of the company. Here, we will not show the process and details of the binomial tree model and only give the result of it. With same data of the company, the price of each share of the stock given by the binomial tree model is \$15.9 [12]. Obviously, the Black-Scholes model with the application of SDE can reach a high accuracy comparing with traditional method. The advantage of it lies in that it can be used to price the stock continuously. In other words, it can record the developing process of the stock price of a company.

4. Prospects

The community's awareness of and inclination toward random forest models has considerably increased in recent years. It can be readily applied to practically any data science issue, enabling users to obtain the initial benchmark findings quickly and effectively. Random forest has consistently demonstrated remarkable power in a variety of situations while also being incredibly practical and convenient. The random forest method is applicable to a wide range of industries, including finance, the stock market, medical, and e-commerce. It is used in the banking industry to identify clients who use banking services more frequently than average clients and to collect on their debts on schedule. Additionally, it will be used to identify clients who want to scam banks. It may be used in the medical profession to decide on the ideal medicine formulation and to examine a patient's medical history to find disorders. Random forest may be used in e-commerce to ascertain whether people genuinely enjoy a product.

SDE and the Black-Scholes model, in the meantime, also play an important role in real life. Statistics has always been important in financial field and still will be as the database become larger and more comprehensive. In the Black-Scholes model, more data in history will be collected to give more accurate approximation of the coefficients in the model. In fact, the volatility, which is one of the key factors in the Black-Scholes model, is certain to be optimized with more detailed data of a company and the market. On this basis, we will be able to predict the market price of a certain product more accurately and thus expect a higher profit at a relatively low risk. In this way, the companies can price their stock in a reasonable way and the buyers can make decisions to help them make profit in the stock market. Apart from the application in the stock market, SDE is also expected to be useful in other fields concerning with randomness.

5. Conclusion

In summary, this study discusses the implementation of statistical tools with two specific models, random forest and the Black-Scholes model. Random forest models can help us classify and predict. From the above, we can see that there is a positive correlation between each attribute and legendary. Moreover, we can see that the attributes that can make Pokémon legendary are special attack and speed. Due to the limited data, many of the attributes of Pokémon are not included in the data set, so we can't analyze all the actual attributes of Pokémon. Based on specific data analysis, this example simply and intuitively shows the practical application ability of random forest, and shows its convenience and accuracy for us.

Meanwhile, SDE is also widely applied in statistic models. We have investigated its features and relate it to the application in the Black-Scholes model. In the first part of the research, the relationship between the coefficients and the trajectory of SDE is studied, which provides the basic theory which is important in the application part. In the second part, we use the Black-Scholes model to predict the market price mainly with two parameters, the volatility, and the profit rate. By estimating these two parameters, one can use the Black-Scholes model to predict the market price, and the result turns out to be reasonable. Nevertheless, limitations do exist in our study. This study only considers the situation in which the parameters are constant for sake of simplicity, while both parameters are changing every second in real life circumstances. Therefore, the model can be optimized if we can use another function of time to represent the parameters. Nevertheless, this research still plays an important role. It offers a basic idea on the way to predict the market prices in a scientific way with the help of SDE and Black-Scholes model and can be instructive for people to make their decisions in stock market.

References

[1] Farnaaz N, Jabbar M A. Random forest modeling for network intrusion detection system. Procedia Computer Science, 2016, 89: 213-217.

[2] Belgiu M, Drăguţ L. Random forest in remote sensing: A review of applications and future directions. ISPRS journal of photogrammetry and remote sensing, 2016, 114: 24-31.

[3] Qi Y. Random forest for bioinformatics//Ensemble machine learning. Springer, Boston, MA, 2012: 307-323.

[4] Nualart D, Pardoux É. Stochastic calculus with anticipating integrands. Probability theory and related fields, 1988, 78(4): 535-581.

[5] Scholes M, Black F. The pricing of options and corporate liabilities. Journal of Political Economy, 1973, 81(3): 637-654.

[6] MacBeth J D, Merville L J. An empirical examination of the Black-Scholes call option pricing model. The journal of finance, 1979, 34(5): 1173-1186.

[7] Clark K. Tiers of Competitive Pokemon Battles. 2020.

[8] Navas D, Donohue D. Competitive Pokémon Usage Tier Classification. Advances in Data Science and Information Engineering. Springer, Cham, 2021: 599-607.

[9] Dunbar S R. Mathematical Modeling in Economics and Finance: Probability, Stochastic Processes, and Differential Equations. American Mathematical Soc., 2019.

[10] Saratha S R, Sai Sundara Krishnan G, Bagyalakshmi M, et al. Solving Black–Scholes equations using fractional generalized homotopy analysis method. Computational and Applied Mathematics, 2020, 39(4): 1-35.

[11] Abid Ali Awan. Pokémon Legendary data. 2020. Retrieved from: https://www.kaggle.com/datasets/kingabzpro/pokmon-legendary-data.

[12] Jackwerth J C. Generalized binomial trees. Journal of Derivatives, 1996, 5(2): 7-17.