Predicting IPO initial returns using random forest

Abstract:

Empirical analyses of IPO initial returns are heavily dependent on linear regression models. However, these models can be inefficient due to its sensitivity to outliers which are common in IPO data. In this study, the machine learning method random forest is introduced to deal with the issues the linear regression cannot solve. The random forest is used to predict initial returns of IPOs issued on Borsa Istanbul. The prediction accuracy of the random forest is then tested against methods of robust regression. The prediction results show that random forest has by far outperformed other methods in every category of the comparison. The variable importance measure shows that the IPO proceeds and IPO volume are the most important predictors of IPO initial returns. The results also show that the variables that act as potential proxies for ex-ante uncertainty are more important than variables that are proxies for information asymmetry.

Keywords: Random forest, initial public offerings, initial returns, underpricing, prediction.

JEL classification: G12, G30, G39

1. Introduction:

IPOs initial returns, often referred to as IPO underpricing in the literature, is one of the most renowned market anomalies and has been documented in many markets. As early as the 1970s, researchers such as Ibbotson (1975) observed that the initial performance of IPOs was exceptionally high, Stoll and Curley (1970) as well noticed a remarkable price appreciation of equity offerings between the initial offering date and the first market date. In the following years, IPOs underpricing has been taken seriously and widely discussed in the literature of finance. Loughran et al. (1994) document the occurrence of this phenomenon in 25 markets around the globe. Similarly, Ritter and Welch (2002) find that the offer prices of IPOs issued by US companies were underpriced by an average of 16 percent, this figure then jumped to an extreme level during the internet bubble. Consistent with the global evidence, the IPOs underpricing has been found highly significant in the emerging markets (Huang et al., 2016; Alanazi& Al-Zoubi, 2015; Chang et al., 2008; Kiymaz, 2000). Empirical evidence, however, shows that the level of IPO underpricing differs considerably among countries. Based on data compiled by Ritter (2015) the emerging markets have much higher IPO underpricing ratios compared to developed markets[1]. Engelen and Essen (2010) examined IPO data of 21 countries and found a variation of 10% in the level of underpricing between countries.

Explaining this anomaly has been a prominent focus of academic researchers. Although the notion of IPO underpricing may seem straightforward, practically the process of IPO is characterized by the complexity of determining the offer price. This complexity arises from the potential conflict of interests among the participants in the IPO process. For this reason, there has been little consensus regarding whether the IPO underpricing is a desirable or undesirable outcome of the IPO process. For instance, Dalton et al. (2003) find that in most of the cases the underwriters are not acting in the best of their clients i.e. the IPO firms, but rather favoring the recipients of the IPO shares whom often end up receiving most of the IPO proceeds. On the other hand, Beatty and Ritter (1986) state that excessive underpricing of the IPO by the underwriters would be appealing to the uninformed investors, but it would not be so to the IPO firms. On the contrary, a higher offer price would benefit the IPO firms but discourages the uninformed investors from buying the IPO. Therefore, the underwriters will seek some optimal level of

[1]Updated global IPO underpricing information can be found on Jay Ritter's website: https://site.warrington.ufl.edu/ritter/files/2015/05/Initial-Public-Offerings-International-Insights-2015-05-21.pdf

underpricing that satisfies both sides. Loughran and Ritter (2002) in efforts to understand why the issuers are satisfied leaving a large amount of money on the table by letting the underwriters set a low offer price, they examined the covariance between the issuers' capital sacrifice and their overall wealth after listing. Loughran and Ritter found that the issuers attain larger wealth gain on the retained shares from a price jump. Another aspect of IPO underpricing complexity lies in the difficulty to determine the factors that lead to underpricing. In this respect, a considerable amount of theoretical explanations has been developed to rationalize the anomaly of IPO underpricing. Welch and Ritter (2002) categorize the theories of IPO underpricing into asymmetric and symmetric information. The explanations based on asymmetric information theories have been widely supported and followed in the literature of IPO underpricing. Symmetric information theories, on the other hand, have not been widely accepted as the primary determinant of underpricing. One explanation that falls under this category is the Tinic (1988) argument which suggests that the IPO is intentionally underpriced by the issuers to reduce their legal liability. Welch and Ritter add another category to IPO underpricing theories which focuses on the allocation bias of IPO shares among the investors. Most of these theories have been subjected to rigorous empirical testing using firm-specific and market specific-factors. The empirical evidence presented in the literature is notably in favor of the asymmetric information theory.

In this study, a machine learning algorithm is employed to predict IPOs initial returns. Machine learning is a subset of data science that learn when exposed to a dataset, but the dataset has to be large enough to sustain the learning process, financial data though is considered small to medium compared to other fields where machine learning models are applied to larger data. However, financial data is also featured with noise and might be heavily skewed in some cases, for such issues the nonlinear techniques are the most appropriate. Fortunately, machine learning algorithms fit perfectly for this purpose. In addition, some of the machine learning algorithms have already been used in many financial applications, especially in risk management and forecasting the future stock returns, and have outperformed the classical financial methods (see e.g., Desai and Bharati, 1998; Thawornwong et al., 2003; Maciel and Ballini, 2010; Hanias et al., 2012). A review by Tkac and Verner (2016) shows that artificial neural network applications have made significant inroads in finance over the last two decades. According to this review, the majority of these applications are found in studies related to predicting financial distress and

bankruptcy and forecasting shares and bonds performance. As for predicting IPO initial returns, the linear regression methods are still the dominant approach. However, there have been significant efforts to analyze IPO returns using a variety of machine learning and computational intelligence techniques. To name a few, Luque et al. (2012) focus mainly on the offering characteristics to predict IPO returns using a genetic algorithm. Huang et al. (2012) also apply a genetic-based algorithm on the IPO's fundamental variables to select the potentially high-growth stocks. The artificial neural networks (ANN) and support vector machine (SVM) have been used to predict IPO initial returns in studies such as Mitsdorffer and Diederich (2008) and Bastı et al. (2014). The fuzzy neural network (FNN) an advanced intelligence system of ANN has been applied recently by Wang et al. (2018) to predict the underpricing of a large sample of U.S. IPOs. Robertson et al. (1998) construct an OLS regression and neural network models to predict the first day returns of IPOs, the empirical findings of their study show that the predictions produced by neural network models were better than predictions produced by OLS regression. The same approach was followed by Reber et al. (2005) to predict IPO initial returns. Their results, however, showed a slight advantage of neural network models over OLS regression. More recently, the random forest, a powerful and well-known machine learning algorithm, has been used by Quintana et al. (2017) to predict the IPO underpricing.

2. Literature review:

Ritter and Welch (2002) state that the theories explaining the anomaly of IPO underpricing follow two lines, the first line is based on the information asymmetry problem among the participants in the IPO transactions. The winner's curse model of Rock (1986), the first model to underline the asymmetric information, contends that some investors are better informed about the market than other investors. The informed investors have the edge over other investors as they only subscribe to shares of an attractive IPO while the uninformed investors subscribe to every IPO. This means that attractive shares will be oversubscribed and dominated by informed investors. The uninformed investors, on the other hand, will receive a small proportion of the attractive IPO shares and the full supply of unattractive IPOs. As a result, they end up receiving expected returns below the average underpricing or even negative returns. The uninformed investors then would react by ceasing bidding for any IPO shares in the future. Therefore, Rock (1986) assumes that the underpricing is pre-market determined by the issuers in order to increase the demand and reduce the effect of the information asymmetry between the informed and the

uninformed investors. Beatty and Ritter (1986) conceptually extend the winner's curse model with a different division between informed and uninformed investors, this model assumes that the underwriters are better informed about the market than the issuing corporations. This, in turn, allows them to set the offer price more accurately. The signaling model developed by Welch (1989), Allen and Faulhaber (1989) and Grinblatt and Hwang (1989) also complements the argument of information asymmetry by assuming that the issuing corporations are better informed about the intrinsic value of their companies than the outside investors.

The other line of IPO underpricing theories discusses the allocation bias in the shares of IPO. It argues that the underwriters may use their discretionary power to purposively underprice and diffusely allocate the underpriced IPO shares to favored clients. The underwriters, in turn, can boost their profits through the trading commissions they acquire from the clients.

Despite the myriad of theories put forth by the researchers to explain the anomaly of IPO underpricing, the underlying causes of this phenomenon are still debated. As pointed out by Ljungqvist (2007), the asymmetric information theories are the most supported by empirical evidence, but according to Ritter and Welch (2002), it is unlikely to be the primary determinant of fluctuations in IPO activity and underpricing. In addition, most of the theories hold well in the developed markets, yet they are less successful in explaining extremely high underpricing of IPOs in emerging markets. In China, for example, Titan (2011) and Gao (2010) find no significant relationship between the IPOs initial returns and the proxies of asymmetric information. The studies on the Indian IPOs find mixed evidence. Using IPO data for the period 2005-2012, Chhabra et al. (2017) find the variables that signal information are highly significant and companies with high information disclosure experience less underpricing. On the other hand, Chhabra et al. (2017) find the informational variables less effective in explaining the IPO underpricing.

Table 1
Studies of IPO underpricing from developed and emerging markets.

Authors	Dataset	Variables	The main findings
Pande and Vaidyanathan (2007)	IPOs issued on NSE in the period between March 2004 and October 2006	Dummy variable for market demand, listing delay, marketing expenditures, issue size, market change on the day of listing.	Market demand and market percentage change are the main drivers of IPO underpricing.
Wadhwa (2014)	The analysis was performed on the dataset of 92 IPOs issued between 2009-2011 on NSE	Offer price, listing delay, a dummy variable for market demand, the reputation of the leading underwriter, issue size, IPO grade, firm age, internal risk factors, equity retained, a dummy variable for private and government issues	Underpricing was found to be positively related to the offer price and listing delay.
Lin and Hsu (2008)	The newly listed firms on Taiwan stock exchange and Hong stock exchange for the period between January 1999 and June 2004	Market adjusted returns, Allotment ratio, Debt Ratio, IPO proceeds, trading volume, market cap dummy, sectors' dummy	The evidence does not support the ex-ante uncertainty hypothesis in both markets, asymmetric information measures are the most consistent determinant of underpricing in both stock markets
Ekkayokkaya and Pengniti (2012)	A sample consists of 463 IPOs made during the period 1990–2007, and subsamples for the pre-reform, transitional and post-reform periods	Market co-movement, lagged underpricing, pre-IPO market returns, pre-IPO market volatility, stock market cap. to GDP, aftermarket volatility, issue size to industry market cap., elapsed time, dummy variable for governance reforms, use of proceeds disclosure, ownership control retention.	The study documents a significant reduction in IPO underpricing following major governance reforms. The listing activities show improvement after governance reforms and the frequency of use of proceeds disclosures have reduced the ex-ante uncertainty.
Satta (2017)	IPOs issued on international stock exchanges by firms operating in the port industry in the 2001–2015	Firm age, firm size, core business, number of underwriters, the reputation of the lead underwriter, percentage of equity offered, variable represents the historical background of the port industry.	The findings support the validity of the timing and signaling hypothesis, the age of the issuer is found to moderate IPO underpricing, IPOs issued by port companies suffer higher levels of underpricing.
Peng and Wang (2007)	A sample consists of 647 IPOs issued in Taiwan stock exchange in the period 1996-2003.	Proxy for the flotation method, the return of the market index, the standard deviation of market index returns, the ratio shares offered divided by the number of subscriptions, the offer size, sale revenue, index of earnings management, underwriter's reputation, the auditor reputation, dummy variable for electronic firms, dummy variable for firms listed in the OTC.	Ex-ante uncertainty and asymmetric information play a significant role in IPO underpricing, the auction flotation method reduces IPO underpricing.
Tian (2011)	Dataset consists of 1377 IPOs listed on the Shanghai and Shenzhen Stock Exchanges between 1992 and 2004.	The pricing cap, issuing size, allocation rate, total assets, Firm age, listing delay, size of the government shareholding, size of block shareholding, size of employee shareholding, size of managerial shareholding.	Financial regulations account for more than half of the severe underpricing, investment risks also contribute to severe underpricing, asymmetric information is far from being the major determinant of underpricing.
Jewartowski and Lizińska (2012)	IPOs issued on the Warsaw Stock Exchange in the period 1998-2008.	Firm size, privatization, return on equity, the average of underpricing over a ninety-day window prior to the IPO date, market condition, market-to-book value, early return volatility.	The study documents a strong effect of early aftermarket volatility, issuer's size, growth opportunities, and profitability before the offering on IPO initial returns.
Marshal (2004)	A total of 532 IPOs listed on US stock markets in the period 1993-1995	The rank of the lead underwriter, the auditor reputation, percentage of retained shares, percentage of proceeds allocated for usages outside general corporate or working capital, offer size, percentage of venture capital or corporate funding, debt capacity, IPO volume at the industry level, measures of firm's financial risk	Financially risky firms with few alternative financial sources have higher underpricing, the large supply of IPOs in a particular industry leads to higher initial-day returns.
Falconieri et al. (2009)	IPOs listed on the AMEX, NASDAQ, or NYSE during the period 1993-1998	The standard deviation of return over the first twenty days after listing, the offer size, proxy for hot issues, firm age, a dummy variable for tech and internet companies, the rank of the lead underwriter, first day's trading volume, proxies for ex-post value uncertainty.	Including proxies for ex-post value uncertainty beside ex-ante uncertainty improves the explanatory power of the model.
Vong and Trigueiros (2010)	All the new offerings listed on the Hong Kong Exchange over the period 1994–2005	The subscription rate, the offer size, firm size, the offer price, standard deviation of returns over the first ten days after listing, the market share of the underwriter, IPO volume.	The reputation of underwriters helps to reduce the underpricing, the informed demand hypothesis of Rock (1986) found to be significant under specific conditions.

Note: Except for the study of Peng and Wang (2007) which uses stochastic frontier model, all the other reported studies use linear regression.

3. Data and model selection:

3.1. Data:

The data set used in this study consists of first-day trading returns of 127 public offerings listed on Borsa Istanbul (BIST) during the period 1998-2018. In this period, a total of 217 firms went public raising around 17.8$ billions of capital. As it is shown in table 2, IPO activities peaked in 2000 with 35 deals then dramatically dropped to one deal in the next year, this dramatic decline was due to the liquidity crisis in 2001 and the depreciation of the Turkish Lira by 50% against the dollar in a short period of time. In the following years, especially since 2004, the global favorable monetary conditions spurred economic growth and sectoral developments in many emerging markets. In this period, Turkey saw an unprecedented scale of privatization. The IPO market, in turn, experienced a significant uptick in terms of the realized proceeds. It may be noted that the period 2004-2007 was the most successful period for IPOs thanks to the large scale of privatization sales, a total of 44 firms went public in this period raising around 6.4$ billions of revenues. The period of global recession showed a stark reduction of the number of IPOs with only three listings on the BIST in the years 2008 and 2009 combined. This reduction in IPO activity was largely driven by the rapid decline in prices on BIST following the breakout of the global recession in the third quarter of 2008. The IPO activities bounced back in 2010 and maintained its positive trend up until 2012, albeit with a reduced total amount of capital raised. The IPO market in Turkey begun to slow down again in 2012 and has not been so active in the following years. Throughout the period of the study, we find that 127 offerings out of the total offerings were underpriced, the rest of the offerings which account for around 41.5% of the total offerings were either overpriced or correctly priced.

The IPO data used in the empirical analysis is obtained from Borsa Istanbul. The data of the companies' financial operating history prior to IPO was extracted from the financial statements archive of the publicly traded companies provided by Borsa Istanbul and the public disclosure platform (kap.org.tr). Information about the foundation dates was retrieved from the companies' yearbooks provided by Borsa Istanbul.

Table 2
IPO activities on Borsa Istanbul during the period 1998-2018.

Year	Number of IPO deals	IPO proceeds (millions of USD)	Percent (%)
1998	20	383.35	2.15
1999	10	90.72	0.51
2000	35	2806.22	15.76
2001	1	0.24	0.00
2002	4	56.47	0.32
2003	2	11.25	0.06
2004	12	482.58	2.71
2005	9	1743.96	9.80
2006	15	930.50	5.23
2007	9	3298.31	18.53
2008	2	1876.92	10.54
2009	1	6.91	0.04
2010	22	2104.02	11.82
2011	25	826.49	4.64
2012	16	297.08	1.67
2013	8	721.65	4.05
2014	9	308.87	1.73
2015	3	23.21	0.13
2016	2	117.14	0.66
2017	3	351.76	1.98
2018	9	1366.44	7.67
Total	217	17804.1	100

Source: Borsa Istanbul (BIST).

3.2. Model selection:

To predict the IPO first-day returns, we first select the most commonly used measure of underpricing proposed by Ibbotson and Jaffe (1975) and Ibbotson et al. (1988, 1994), which is expressed as

$$IR_{it} = (CP_{it} - OP_{it})/OP_{it}$$

where CP_{it} is the closing price of the first trading day; $OP_{it}$is the firm's offer price i at time t; IR_{it} is the IPO's initial return of the firm i at time t.

Secondly, the IPO initial return is regressed over a set of variables that have been theoretically and empirically linked to underpricing in the preceding literature. The variables used in the empirical model can be classified into company characteristics, offering characteristics, and market sentiment indicators.

Company characteristics: This category consists of the variables firm age, firm size, net income and returns on assets (ROA) to proxy for the firm-specific risk factors. The firm age and firm size are commonly used in the literature as a proxy to measure uncertainty and information asymmetry as well. Firms established long before listing should have more information available to the public which reduces the information asymmetry among the issuer, the underwriters, and the investors. Ritter (1984) argues that the level of underpricing is positively related to ex-ante uncertainty about the value of the firm. The level of underpricing, therefore, is negatively associated with the company's age prior to listing. This relationship has been empirically confirmed in many studies (Ange & Brau, 2002; Loughran & Ritter, 2004; Chahine, 2008). Similarly, larger firms, as compared to small firms, are perceived as less risky because it displays less uncertainty about its value. The large companies also have more public disclosure which decreases information asymmetry, winner's curse theory of Rock (1986) and information asymmetry theory of Beatty and Ritter (1986) both suggest that greater information asymmetry is always associated with bigger underpricing. Net income and ROA are also included to proxy for information asymmetry and to evaluate the firm's quality and performance prior to IPO, Lowry and Shu (2002) state that firms that experience stronger operating performance prior to the IPO are subject to less uncertainty.

Offering characteristics: The variables selected under this category are the IPO volume, the IPO rate, the offer price, and IPO proceeds[2]. The offer price, IPO proceeds, and volume can be regarded as an indicator of uncertainty. Daily et al. (2003) suggest that the highly-priced IPOs are characterized by lower uncertainty regarding the future performance of the firm. In contrast, Ibbotson (1988) finds firms that offer a low price have a high level of uncertainty, and that their offered equities can be subjected to speculative trading. The size of the IPO often measured by the proceeds is supposed to have a strong impact on underpricing, Clarkson (1994) find the IPO size to be an effective proxy for ex-ante uncertainty, whereas How et al. (1995) report a significant negative relationship between the size and underpricing. As for the IPO rate, which represents the percentage of shares offered, the signaling theory suggests that this factor conveys information about the quality of the firm (Leland and Pyle, 1977). Consequently, the higher

[2]To remove the inflation effect that may distort the results, the IPO proceeds specifically have been taken as US dollars in most of the studies on Turkish companies' IPOs such as Bastı et al. (2015), Durukan (2006),Yüksel and Yüksel (2006), Durukan (2002) and Kiymaz (2000).

percentage of shares offered, the lower underpricing. In addition to these variables, a dummy variable is included to indicate for the foreign investors' participation in the IPO process.

Market sentiments: The 30 days and 60 days of market performance prior to IPO date are commonly used as an indicator for market sentiments. The cyclical behavior hypothesis argues that IPOs realized during the hot market are heavily underpriced compared to those realized in periods of cold market. Ritter (1984) documented the existence of these type of behaviors in US markets during the 1980s. In addition, Hanley (1993) reports a positive relationship between IPO's initial return and market index returns prior to IPO. Due to the existence of a short-term momentum effect in Borsa Istanbul as evidenced by Ejaz and Polak (2015), the two measures are considered in the empirical analysis. Therefore, two different empirical models are developed to study the effect of each measure separately.

Table 3
Variables definitions.

Dependent variable	IR	Fist-day IPO return
Firm characteristics	FA	Firm age prior to the listing date
	FS	Firm size (assets)
	ROA	Return on assets
	NI	Net income
Offering characteristics	IPOV	Number of shares offered
	IPOR	IPO rate (the proportion of shares offered)
	OP	The offer price
	Pd	The IPO proceeds (in USD)
	D20	Dummy variable equals 1 if the foreign investors' purchase of IPO exceeds 20%, and 0 otherwise.
Market sentiments	MP30	30 days market return prior to IPO
	MP60	60 days market return prior to IPO

4. Research methodology:

The empirical analyses carried out to explain and predict the IPO initial returns are often based on linear regression models. There are also cases where non-linear models such as logistic regression are implemented. The use of machine-based methods has been on the rise recently. In this study, the random forest -one of the most popular machine methods in both classification and regression- is used to predict the returns of IPOs issued on Borsa Istanbul in the period between 1998-2018. In addition, since the random forest is a novel technique to the IPO literature, its predictive accuracy is compared to some of the well-known robust regression methods. To the best of the authors' knowledge, there is only one study conducted by Quintana

et al. (2017) which uses the random forest to predict the IPO returns. According to Quintana et al. (2017), the random forest has some unique features over other tree-based techniques which make it potentially suitable for predicting IPO returns. Predicting IPO initial return has been a challenging task due to the involvement of a large number of determinants with very different explanatory power and the presence of outliers. In this regard, the random forest with its ability to combine weak and strong variables and its robustness to outliers can be a very useful tool for this task. In general, the machine learning algorithms and particularly the random forest work effectively on large data. Therefore, sectoral segmentation of the IPOs which is a common practice in the IPO literature would significantly shrink the data sample and ultimately leads to poor results. On account of this, sectoral segmentation of the IPOs is avoided to ensure better results. The main purpose of this study is to extend the work of Quintana et al. (2017) to other markets and provide further supporting evidence to the advantage of using the random forest in predicting IPO initial returns.

4.1. Random forest:

Random forest, developed by Breiman (2001), is an ensemble learning method in which multiple decision trees are constructed and merged together to get a more accurate and stable prediction or classification. The trees in the random forest are drawn from the original sample using bootstrap resampling, and each tree is grown using a randomized subset of features. The procedures to produce the random forest of regression trees are explained below.

Let's assume we have the dataset $D = \{(x_1, y_1) \dots \dots (x_n, y_n)\}$ and the aim is to find the function $f: X \rightarrow Y$ where X is the inputs and Y is the produced outputs. Furthermore, let M be the number of inputs.

1. Random forest randomly selects n observations from D with replacement to form a bootstrap sample.

2. Each tree is grown using a subset of m features from the overall M features. For regression, it is recommended to set the subset of features at $M/3$.Then at each node, m features are selected at random and the best performing split among the m features is selected according to the impurity measure (Gini impurity).

3. The trees are grown to a maximum depth without pruning.

Growing trees without pruning and selecting the best features split at each node allow the random forest to maintain prediction strength. In addition, the random selection of features reduces the correlation between the trees. Unlike other tree-based techniques, the random forest is immune to overfitting as more trees are added to the forest. The overall predictions are produced by taking the average of the predictions of the individual trees in the forest. Random forests do not only generalize the predictions of trees over the entire sample, but also provide variable importance measure using the out-of-bag sample. The main idea is eliminating the dependence of the predictor with the response variable by permuting its values across all trees, then the loss of prediction accuracy of the forest is estimated, high loss implies high importance of the predictor and vice versa. It should be noted that random forest is frequently implemented with K-folds cross-validation when accurate assessment against other machine methods is required. However, such a procedure may not be necessary since the unbiased estimate of error can be estimated internally in the random forest.

4.2. Robust regression methods:

The ordinary least squares method is known to be sensitive to outlier points. In robust regression, the influence of the outliers on the fitted regression line is reduced using weight function. This has the additional advantage of making outliers stand out more strongly against the line. There are many weighting functions proposed in the literature. In this study, three robust regression methods are used namely the iteratively reweighted least-squares (IRLS) algorithm, the least median squares (LMS), and the least trimmed squares (LTS). In the IRLS algorithm, the outlier points are weighted using Huber psi and Tukey's bisquare psi functions.

4.2.1. The iteratively reweighted least-squares (IRLS):

The method of iteratively reweighted least squares consists of an underlying weighted least squares fit that is placed inside an iteration loop. When a loop is passed at each iteration, a least-squares fit is carried out using a set of weights, weight is assigned to each observation. Moreover, the weights are derived from the current residual and are updated from iteration to iteration. Since the new weights are derived from the residuals, the iteration process goes on as the residuals keep changing, then the process terminates when the residuals remain unchanged over two passes. The IRLS heavily depends on the weighting functions (see Heiberger and Becker. (1992) for more details about the most commonly used weighting functions).

4.2.2. The least median squares (LMS) and least trimmed squares (LTS):

Linear least squares estimator minimizes the sum of squared residuals to find the parameters that best fit a set of data points. The least median squares estimator replaces the sum of squared residual with the median of squared residuals. As stated by Rousseeuw (1984), the creator of the technique, the resulting estimator from this process can resist the effect of nearly 50% of contamination in the data. Rousseeuw later introduced the least trimmed squares (LTS) to improve the asymptotic efficiency of LMS. This method consists of finding a subset of data points whose deletion from the data set would lead to the regression with the smallest residual sum of squares. The idea is to mitigate the influence of outlier points by minimizing the sum of the smallest squared residuals rather than the complete sum of squares. This is done by ordering the squared residuals from smallest to largest, then the number of the smallest squared residuals is determined by specified trimming parameter which also leaves out the percentage of outliers among all the observations, the trimming parameter should at least be more than half of the number of observations. Put differently, the trimming parameter is the threshold that separates the outlier points from the rest of the observations. Therefore, LTS would be less efficient if the trimming parameter is small. However, it should be noted that LMS and LTS have a high percentage of breakdown points, which means that these two methods are insensitive to outliers.

5. Empirical results:

The ability of the random forest to produce accurate predictions of IPO initial returns is the main focus of empirical analysis. To this end, the predictive accuracy of random forest is compared to that of robust regressions in term of mean square errors (MSE) and root mean square errors (RMSE). In addition, the comparison also includes the measures of descriptive statistics of the predictions produced by each method and the actual initial returns. The second part of the analysis discussion is devoted to studying the relative explanatory power of the independent variables.

Table 4
Descriptive statistics.

	Min.	Median	Mean	Max.	Std.dev
IR	0.0032	0.12	0.1314	0.38	0.08024737
FA	0.2849	12.7288	15.4671	72.211	13.89858
FS (in millions)	0.1769	37.55	585	34480	3176.11346
IPOV (in millions)	0.05	3.6	21.520896	625	71.038799
ROA	-0.10143	0.03719	0.07669	0.77646	0.1149453
NI (in millions)	-62.976	0.880541	12.160637	864.259	79.532473
IPOR	0.00345	0.2518	0.30292	0.9907	0.1773995
OP	1	4	10.06	100	14.33584
Pd (in millions $)	0.2427	13.97	89.81	1837	264.401905
MP30	24.0255	0.4077	1.5025	47.3977	11.83767
MP60	-34.049	2.777	2.893	59.782	14.46375

5.1. Predictions results:

To obtain the best predictions from the random forest, the initial value of features split to be used at each node was set at five, and all the other parameters of the trees to be grown were let at default. Then, the value of features split was decreased gradually at each experiment. The best predictions were produced at value three, meaning that all the trees inside the forest were constructed using three random variables. In addition, the same value of features split was given when automated search for the optimal value was implemented, the automated search was executed with the value of features split initially set at two. During the course of the experiments, we noticed that the change in the number of grown trees did not have a major impact on the results, but the change of features split value was crucial in obtaining the best predictions.

Table 5
Prediction errors.

Method	First model		Second model	
	MSE	RMSE	MSE	RMSE
Random forest	0.001156	0.033998	0.001133	0.033669
LMS	0.050222	0.224102	0.045886	0.214211
LTS	0.036997	0.192346	0.013282	0.115249
IRLS-T	0.006011	0.077536	0.005949	0.077135
IRLS-H	0.006012	0.077541	0.006022	0.077603

Note: IRLS-T and IRLS-H refer to iteratively reweighted least squares implemented with Tukey's bisquare psi function and Huber psi function respectively.

Table 5 summarizes the prediction errors measured by mean square and root mean square of errors. MSE and RMSE are measures of the absolute fit of the regression model predictions to the observed values, which also refers to the proximity of predicted values to the observed values and in the same time it indicates for the unexplained variance in the residuals. Therefore, lower values of MSE and RMSE indicate better fit. The random forest as shown table 5 is able to produce far better predictions compared to the other methods, this can be seen more clearly in the plot of the random forest predictions against the actual values in Fig.1. The errors found in the random forest predictions are extremely small in both models, but the predictions of the second model which accounts for the 60 days pre-IPO market performance are slightly better than the predictions of the first model, this means that the market short-term momentum effect is an important factor for the initial performance of IPOs. In fact, all the other models performed better when the 30 days pre-IPO market performance was replaced with the 60 days of pre-IPO market performance. LMS and LTS methods, which have been established to be insensitive to outliers, have performed poorly in both models and they even produced less accurate predictions than the IRLS methods. This was not expected considering the high breakdown point percentage of LMS and LTS, but it should be noted that LMS and LTS do not weigh down the outlier points but rather ignore them. Therefore, it is highly likely that the strong presence of outlier points in the data has far exceeded the resistance level of the methods. The IRLS methods, on the other hand, were able to perform better because it weighs down each observation.

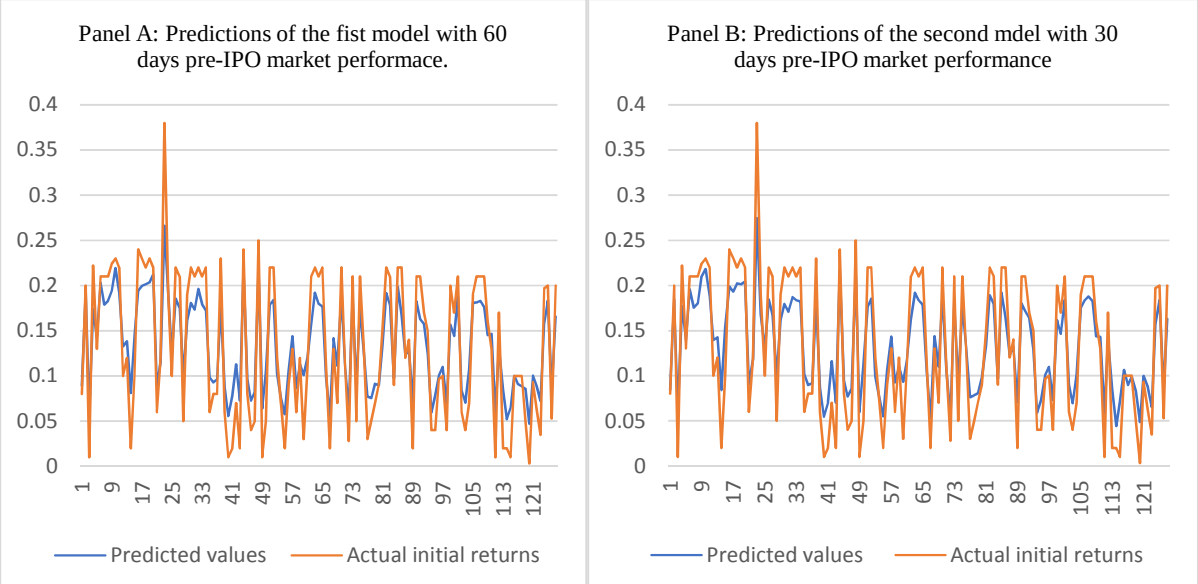

Fig. 1. Random forest predictions vs observed values.

In term of the descriptive statistics of the predictions, random forest performance again has been exceptionally strong, the median and mean predictions by random forest fall close to the mean and the median of the observed values, even the standard deviation of random forest predictions is slightly different than the standard deviation of the observed values. The IRLS predictions in term of the mean and median are acceptable, but the standard deviations of their predictions are far less than the standard deviation of the observed values. The MLS and LTS offered unreliable predictions in each of the descriptive statistics measures. Overall, the descriptive statistics measures make it apparent that the distribution of random forest predictions is relatively identical to the distribution of the observed values.

The IPO underpricing in Turkey has been explored in a number of studies (see e.g., Kiymaz, 2000; Durukan, 2002; Aktas et al., 2003; Yüksel and Yüksel, 2006; Bildik and Yılmaz, 2008; and Ozdemir and Kizildag, 2017). However, all these studies were basically concerned with demonstrating the significance of factors affecting the initial returns rather than making predictions of IPO returns. Moreover, some of these studies focus on a certain aspect of IPO underpricing. For instance, Ozdemir and Kizildag (2017) study the relationship between the franchising activities of IPO candidates and the price of their initial offerings, whereas Yüksel and Yüksel (2006) focus on the effect of the trading volume on underpricing. In addition, the empirical investigations of these studies were carried out using linear regression methods. Meaning that predicting IPO initial returns was beyond the scope of these studies. Therefore, the lack of research on this matter urges the need to explore the subject of underpricing with methods other than classical linear models. In this regard, the random forest can deliver better results on both ends. As it is already shown, the random forest is able to produce accurate predictions for IPO returns. On top of this, variable importance measure, which is statistically equivalent to variable significance in linear regression, can be carried out with the random forest.

Table 6
Descriptive statistics of the predicted values.

	Method of prediction	Median	Mean	Std.dev
First model	Random forest	0.126	0.1321	0.050041
	LMS	0.0736	0.1035	0.171793
	LTS	0.1942	0.165	0.178523
	IRLS-T	0.1303	0.1312	0.020103
	IRLS-H	0.1299	0.1306	0.019427
Second model	Random forest	0.1319	0.1321	0.050117
	LMS	0.1852	0.184	0.192766
	LTS	0.1811	0.1621	0.090621
	IRLS-T	0.1307	0.1311	0.021117
	IRLS-H	0.1291	0.1297	0.019409

5.2. Variable importance:

As previously mentioned, the variable importance is measured by the loss of the model's prediction accuracy when the variable of interest is disassociated from the response variable. The results of this process are represented by Fig. 2. In both models, the variable of IPO proceeds is ranked as the most relevant variable to IPO underpricing followed by the volume of the IPO. Note that introducing the 60 days pre-IPO market performance to the model changed some of the outcomes. The pre-IPO market performance and net income maintained the same rank, but the firm size jumped up two rows and the variable of return on assets fell to the rank five in the second model. In addition, the importance of the offer price decreased while the IPO rate importance increased in the second model. The firm age and the dummy variable signaling for foreign investors' share of IPO are reportedly the least relevant variables. As can be seen in the figure below, the two variables exchange the bottom positions between the two models. The results also suggest that the offerings characteristics represented by IPO proceeds and IPO volume are the main determinants of IPO initial returns, the studies of Durukan (2002), Aktas et al. (2003), Ertuna et al (2003) and Yüksel and Yüksel (2006) all report significant relationship between IPO returns and the IPO proceeds. The IPO rate, the supposed measure of firm quality received an average score in the second model, this variable has been found to be significant in Ertuna et al. (2003) while Kiymaz (2000) reports an insignificant relationship between the IPO rate and IPO returns. The market sentiment is relatively important in both models, Kiymaz (2000) finds that the market trend between IPO date and first trading day of IPO has a significant

impact on IPO returns. Similarly, Yüksel and Yüksel (2006) find the 40 days change in market index prior to IPO to be significant. For company characteristics, firm size and return on assets are the most important for predicting IPO initial returns but not as important as the offering characteristics. The firm's size and age have been reported to be significant in most of the aforementioned studies except in Kiymaz (2000). The variables that act as potential proxies for ex-ante uncertainty such as proceeds, IPO volume, and firm size are highly important in predicting IPO returns, while proxies for information asymmetry such as firm age, return on assets and net income has been less important than proxies for ex-ante uncertainty.

The use of machine learning methods to study IPOs of Turkish companies has been seen previously in a study conducted by Bastı et al. (2014), this study investigated the short-term performance of IPOs using decision trees and support vector machines. To calculate the variable importance Bastı et al. (2014) employed sensitivity analysis, in which the importance scores given by each method were combined and averaged on the weight of the model to obtain the final importance score. In contrast to our study, their findings suggest that the proceeds to be the least relevant variable while the market sentiments had the highest score. However, the sample used in the study of Bastı et al. (2014) consists of the IPOs made by all the companies except investment trusts. In addition, before the execution of empirical analysis the data was screened and the outlier points were cleaned from the sample, such procedure besides being questionable, it may have significantly affected their results. Their findings though were partially similar to the results of this study in the sense that proxies for information asymmetry being less important in predicting IPO returns.

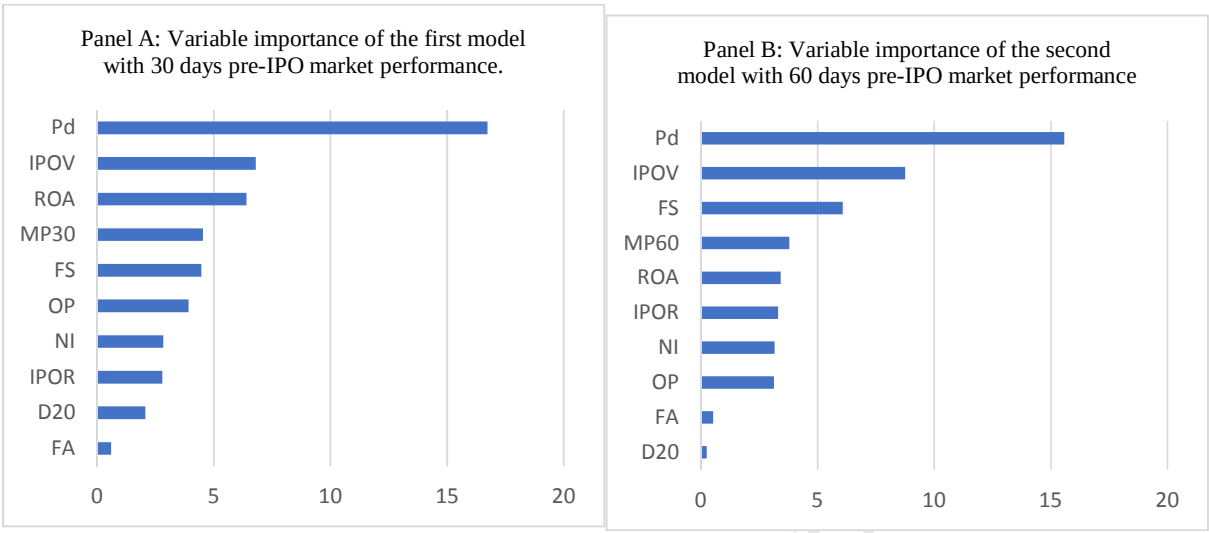

Fig. 2. Out-of-bag variable importance measured by the increase in MSE

Practically, the empirical findings of this study may have further importance in light of the complications involved in the IPO process. These complications, which the bulk of IPO literature hinge on, stem mainly from the determination of IPO price and the post-market IPO performance. In this regard, the use of the random forest to analyze IPO returns may be of particular relevance to the IPO's issuers and investors as both parties are highly concerned with the uncertainty regarding IPO price and performance.

6. Conclusion:

Predicting IPO initial return has been a challenging task due to the involvement of a large number of determinants with very different explanatory power and the presence of outliers. Linear regression models have dominated the empirical investigations in this domain. Linear regressions, however, can be inefficient due to its high sensitivity to outlier points. In this study, the random forest, a powerful machine method, is introduced to deal with the issues linear regression cannot solve. The predictive performance of random forest is tested on a sample of underpriced equity offerings issued on Borsa Istanbul in the period between 1998-2018. Then the results of the random forest are compared to the prediction accuracy of robust regression methods in term of MSE, RMSE, mean, median and standard deviation. Robust regressions are by design less sensitive to data contamination. The LMS uses the median of squared errors instead of the sum while the LTS uses the trimming parameter to exclude the outlier points, in IRLS every point in the data is weighted using weighting functions. The outcomes of the comparison show that random forest has by far the better performance in every category of the

comparison. The random forest predictions for the mean and the median are almost identical to the observed values, while the standard deviation of random forest predictions falls slightly below the standard deviation of the observed values. In fact, random forest predictions are obviously close to the actual initial returns as demonstrated in Fig. 01. The main conclusion to be drawn from this comparison is that random forest has all the potentials to be an ideal alternative to linear regression methods which are commonly implemented to explain IPO returns. In addition, the predictive accuracy of the random forest can potentially boost and facilitate the decision-making process for the IPO participants. In the IPO market where making decisions is probably highly uncertain compared to other stock market activities, IPO issuers, as well as investors, can potentially benefit from the use of the random forest algorithm to reduce the uncertainty regarding the IPO pricing and the post-market IPO performance.

The variable importance of independent variables was also studied using out-of-bag mean square error. The results reveal that the IPO proceeds and IPO volume to be the most important predictors of IPO initial returns. Market sentiments return on assets and firm size were shown to have less explanatory power in predicting IPO initial returns. In general, the results show that the variables that act as potential proxies for ex-ante uncertainty are more important than variables that are proxies for information asymmetry.

References

Alanazi, A. S., & Al-Zoubi, H. A. (2015). Extreme IPO underpricing and the legal environment in wealthy emerging economies. *Journal of Multinational Financial Management, 31*, 83-103.

Allen, F., & Faulhaber, G. R. (1989). Signaling by underpricing in the IPO market. *Journal of Financial Economics, 23*(2), 303-323.

Alok, P., & Vaidyanathan, R. (2009). Determinants of IPO underpricing in national stock exchange of India. *Journal of Applied Finance, 15*(1), 14-30.

Ang, J. S., & Brau, J. C. (2002). Firm transparency and the costs of going public. *The Journal of Financial Research, 25*(1), 2002.

Ataş, R., Karan, M. B., & Aydoğan, K. (2003). Forecasting short run performance of initial public offerings the Istanbul Stock Exchange. *The Journal of Entrepreneurial Finance, 8*(1), 69-85.

Banu, D. (2002). The relationship between IPO returns and factors influencing IPO performance. *Managerial Finance, 28*(2), 18-38.

Bastı, E., Kuzey, C., & Dursun, D. (2015). Analy\ing initial public offerings short-term performance using decision trees and SVMs. *Decision Support Systems, 73*, 15-27.

Beatty, R. P., & Ritter, J. R. (1986). Investment banking, reputation, and the underpricing of initial public offerings. *Journal of Financial Economics, 15*(2), 213-232.

Bildik, R., & Yılmaz, M. K. (2008). The market performance of initial public offerings in the İstanbul stock exchange. *BDDK Bankacılık ve Finansal Piyasalar, 2*(2), 49-75.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32.

Butler, A. W., Keefe, M. O., & Kieschnick, R. (2014). Robust determinants of IPO underpricing and their implications for IPO research. *Journal of Corporate Finance, 27*(C), 367-383.

Chahine, S. (2008). Underpricing versus gross spread: New evidence on the effect of sold shares at the time of IPOs. *Journal of Multinational Financial Management, 18*(2), 2008.

Chang, E., Chen, C., Chi, J., & Young, M. (2008). IPO underpricing in China: New evidence from the primary and secondary markets. *Emerging Markets Review, 9*(1), 1-16.

Chhabra, S., Kiran, R., & Sah, A. N. (2017). Information asymmetry leads to underpricing: Validation through SEM for Indian IPOs. *Program, 51*(2), 116-131.

Chhabra, S., Kiran, R., Sah, A. N., & Sharma, V. (2017). Information and performance optimization: A study of Indian IPOs during 2005-2012. *Program, 51*(4), 458-471.

Clarkson, P. M. (1994). The underpricing of initial public offerings, ex ante uncertainty, and proxy selection. *Accounting & Finance, 34*(2), 67-78.

Daily, C. M., Certo, T. S., Dalton, D. R., & Roengpitya, R. (2003). IPO underpricing: A meta-analysis and research synthesis. *Entrepreneurship: Theory and Practice, 27*(3), 271-295.

Dalton, D. R., Certo, T. S., & Daily, C. M. (2003). Initial public offerings as web of conflicts of interests: An empirical assessment. *Business Ethics Quarterly, 13*(3), 289-314.

Desai, V. S., & Bharati, R. (1998). A comparison of linear regression and neural network methods for predicting excess returns on large stocks. *Annals of Operations Research, 78*(0), 127-163.

Durukan, B. M. (2002). The relationship between IPO returns and factors influencing IPO performance: Case of Istanbul stock exchange. *Managerial Finance, 28*(2), 18-38.

Durukan, B. M. (2006). IPO underpricing and ownership structure: Evidence from Istanbul stock exchange. In G. N. Gregoriou, *Initial Public Offerings: An International Perspective* (pp. 263-278). Butterworth-Heinemann: Elsevier.

Ejaz, A., & Polak, P. (2015). Existence of short term momentum effect and stock market of Turkey. *Investment Management and Financial Innovations, 12*(4), 9-15.

Ekkayokkaya, M., & Pengniti, T. (2012). Governance and IPO underpricing. *Journal of Corporate Finance, 18*(2), 238-253.

Engelen, P. J., & Van Essen, M. (2010). Underpricing of IPOs: Firm-, issue- and country-specific characteristics. *Journal of Banking & Finance, 34*(8), 1958-1969.

Ertuna, B., Ercan, M., & Akgiray, V. (2003). The effect of the issuer-underwriter relationship on IPOs: The case of an emerging market. *Journal of Entrepreneurial Finance and Business Ventures, 8*(3), 43-55.

Falconieri, S., Murphy, A., & Weaver, D. (2009). Underpricing and ex-post value uncertainty. *Financial Management, 38*(2), 285-300.

Grinblatt, M., & Hwang, C. Y. (1989). Signaling and the pricing of New Issues. *The Journal of Finance, 44*(2), 393-420.

Hanias, M., Thalassinos, E. L., & Curtis, P. (2012). Time series prediction with neural networks for the Athens Stock Exchange indicator. *European Research Studies Journal, 15*(2), 23-31.

Hanley, K. W. (1993). The underpricing of initial public offerings and the partial adjustment phenomenon. *Journal of Financial Economics, 34*(2), 231-250.

Heiberger, R. M., & Becker, R. A. (1992). Design of an S function for robust regression using iteratively reweighted least squares. *Journal of Computational and Graphical Statistics, 1*(3), 181-196.

How, J. C., Izan, H. Y., & Monroe, G. S. (1995). Differential information and the underpricing of initial public offerings: Australian evidence. *Accounting and Finance, 35*(1), 87-105.

Huang, C., Chang, C., Kuo, L., Lin, B., Hsieh, T., & Chang, B. (2012). A genetic-search model for first-day returns using IPO fundamentals. *International Conference on Machine Learning and Cybernetics* (pp. 1662-1667). Xian: IEEE.

Huang, S. Y., Lee, C.-H., Pan, L.-H., & Nguyen Thi, B. H. (2016). IPO initial excess return in an emerging market: Evidence from Vietnam's stock exchanges. *Review of Pacific Basin Financial Markets and Policies, 19*(2), 1-23.

Ibbotson, R. G. (1975). Price performance of common stock new issues. *Journal of Financial Economics, 2*(3), 235-272.

Ibbotson, R. G., Sindelar, J. L., & Ritter, J. R. (1988). Initial public offerings. *Journal of Applied Corporate Finance, 1*(2), 37-45.

Ibbotson, R. G., Sindelar, J. L., & Ritter, J. R. (1994). The market's problems with the pricing of initial public offerings. *Journal of Applied Corporate Finance, 7*(1), 66-74.

Jewartowski, T., & Lizinska, J. (2012). Short- and long-term performance of Polish IPOs. *Emerging Markets Finance and Trade, 48*(2), 59-75.

Leland, H. E., & Pyle, D. H. (1977). Informantional asymmetries, financial structure, and financial intermediation. *The Journal of Finance, 32*(2), 371-387.

Lin, C., & Hsu, S. (2008). Determinants of the initial IPO performance: Evidence from Hong Kong and Taiwan. *Applied Financial Economics, 18*(12), 955-963.

Ljungqvist, A. (2007). IPO Underpricing. In E. B. Eckbo, *Handbook of Empirical Corporate Finance* (Vol. 1, pp. 375-422). New York: Elsevier.

Loughran, T., & Ritter, J. (2004). Why has IPO underpricing changed over time? *Financial Management, 33*(3), 5-37.

Loughran, T., Ritter, R. J., & Rydqvist, K. (1994). Initial public offerings: International insights. *Pacific-Basin Finance Journal, 2*(3), 165-199.

Lowry, M., & Shu, S. (2002). Litigation risk and IPO underpricing. *Journal of Financial Economics, 65*(3), 309-335.

Luque, C., Quintana, D., & Isasi, P. (2012). Predicting IPO underpricing with genetic algorithms. *International Journal of Artificial Intelligence, 8*(S12), 133-146.

Maciel, L., & Ballini, R. (2010). Neural networks applied to stock market forecasting: An empirical analysis. *Journal of the Brazilian Neural Network Society, 8*(1), 3-22.

Marshal, B. B. (2004). The effect of firm financial characteristics and the availability of alternative finance on IPO underpricing. *Journal of Economics and Finance, 28*(1), 88-103.

Mitsdorffer, R., & Diederich, J. (2008). Prediction of first-day returns of initial public offering in the US stock market using extraction from support vector machines. In J. Diederich, *Rule Extraction from Support Vector Machines* (Vol. 80, pp. 185-203). Berlin: Springer.

Ozdemir, O., & Kizildag, M. (2017). Does franchising matter on IPO performance? An examination of underpricing post-IPO performance. *International Journal of Contemporary Hospitality Management, 29*(10), 2535-2555.

Peng, Y., & Wang, K. (2007). IPO underpricing and flotation methods in Taiwan-a stochastic frontier approach. *Applied Economics, 39*(21), 2785-2796.

Quintana, D., Saez, Y., & Isasi, P. (2017). Random forest prediction of IPO underpricing. *Applied Science, 6*(7).

Reber, B., Berry, B., & Toms, S. (2005). Predicting mispricing of initial public offerings. *Intelligent Systems in Accounting, Finance and Management, 13*(1), 41-59.

Ritter, J. R. (1984). The "Hot Issue" market of 1980. *The Journal of Business, 57*(2), 215-240.

Ritter, J. R., & Welch, I. (2002). A review of IPO activity, pricing, and allocations. *The Journal of Finance, 57*(4), 1795-1828.

Robertson, S. J., Golden, B. L., Runger, G. C., & Wasil, E. E. (1998). Neural network models for initial public offerings. *Neurocomputing, 18*(3), 165-182.

Rock, K. (1986). Why new issues are underpriced. *Journal of Financial Economics, 15*(2), 187-212.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association, 79*(388), 871-880.

Satta, G. (2017). Initial public offerings in the port industry: Exploring the determinants of underpricing. *Maritime Policy & Management, 44*(8), 1012-1033.

Stoll, H. R., & Curley, A. J. (1970). Small business and the new issues market for equities. *The Journal of Financial and Quantitative Analysis, 5*(3), 309-322.

Thawornwong, S., Enke, D., & Dagli, C. (2003). Neural networks as a decision maker for stock trading: A technical analysis approach. *International Journal of Smart Engineering System Design, 5*(4), 313-325.

Tian, L. (2011). Regulatory underpricing: Determinants of Chinese extreme IPO returns. *Journal of Empirical Finance, 18*(1), 78-90.

Tkac, M., & Verner, R. (2016). Artificial neural networks in business: Two decades of research. *Applied Soft Computing, 38*, 788-804.

Vong, A. P., & Trigueiros, D. (2010). The short-run price performance of initial public offerings in Hong Kong: New evidence. *Global Finance Journal, 21*(3), 253-261.

Wadhwa, B. (2014). Insights into the IPO underpricing for listing on National stock exchange. *Journal of Business Thought, 5*, 38-58.

Wang, D., Qian, X., Quek, C., Tan, A., Miao, C., Zhang, X., . . . Zhou, Y. (2018). An interpretable neural fuzzy inference for predictions of underpricing in initial public offerings. *30*, 102-117.

Welch, I. (1989). Seasoned offerings, limitation cost and the underpricing of initial public offerings. *The Journal of Finance, 44*(2), 421-449.

Yüksel, A., & Yüksel, A. (2006). The link between IPO underpricing and trading volume: Evidence from the Istanbul Stock Exchange. *Journal of Entrepreneurial Finance and Business Ventures, 11*(3), 57-78.