利用机器学习和情感分析算法预测股票价格——以特斯拉公司为例

贾景翔
人民大学附属中学

日期： 2021 年 9 月 23 日

摘　　要

本文的主要目的是分析如何结合机器学习方法和情感分析对股票价格进行准确预测。为了实现这一点， 我们以特斯拉公司 2020 年全年的股票数据和推特用户对特斯拉公司的推文数据为例， 量化了网民对该公司的情绪， 运用多个机器学习模型对未来股票价格进行预测。经过研究， 得出以下结论： 基于历史股价数据可以有效预测未来股价； 不同机器学习算法对股票价格预测效果不同； 增加情绪数据可以增强深度学习（LSTM) 和线性回归模型的预测准确率，降低 K 近邻模型、支持向量机模型和决策树模型的准确度。

关键词： 股价预测, 机器学习, 情感分析, K 近邻, 决策树，支持向量机， LSTM

1　导言

1.1　研究背景

股票是一种重要的投资选择， 股票价格的变化也因此受到人们的关注。同时对股票市场的研究也是学术界研究的热点。传统金融学认为， 投资是理性的， 可以根据获得的信心做出最优的决策， 股票价格一般是由股票的实际价值决定的。股票价格受到市场供给和需求量方面的影响， 而两者之间会由于信息的不对称而影响股票价格的波动。用过研究证券在市场上供求关系的基本因素， 依靠经济学、金融学、财务管理学和投资学等基本原理， 判断证券的合理价位并提出相应投资决策意见的方法称之为基本面法 [1]。

传统的基本面分析模型中， 最早的成果是三因子模型。该模型认为， 一个投资组合（包括单个股票） 的超额回报率可由它对 3 个因子（市场投资组合、市值因子、账面市值比） 的暴露来解释。在三因子模型的基础上， 有学者引入了动量因子， 构造了四因子模型， 它对于基金绩效的解释能力较前者有了很大的提高。四因子模型把基金收益描述为在市场因素、规模因素、价值因素与动量因素共同作用下的结果。后续又有学者在此基础上扩展出五因子模型等。上述一类分析方法称为基本分析法。基本分析法能够很好地反应股票的内在价值， 因而在较长周期中，能够对股票的走势起到很好的预测作用。但基本分析法的缺点是对价格预测反应比较迟钝， 不适合短期预测。

技术指标反映市场某一方面的信息， 它是通过对基本的股票数据使用一些数学公式运算得到的， 可以反映数据的更为抽象的特点， 可以通过图、表和数值进行展现。基于技术指标的分析

方法称为技术分析法， 它忽略了市场的因素和经济定律， 只注重市场的行为， 其包含 3 大假设：1）市场涵盖一切信息， 只分析市场信息可以使分析更为便捷； 2）证券价格沿趋势变动， 其运动方向由供求关系决定； 3）价格的运行方式汪汪会与历史重合。技术分析包括多种方法， 大致分为 K 线理论、波浪理论、技术指标、循环周期、切线理论、形态理论等。市场上的技术指标很多， 最常用的使相对强弱指标（RSI)、简单移动平均（MA）、指数平滑异同平均（MACD)、随机指标（KDJ) 等。基于技术指标， 研究者尝试使用多种建模方法预测股票价格， 如灰色预测模型、马尔可夫链以及各种时序模型（ARIMA 等） [2-4]。近年来， 机器学习技术发展迅猛， 涌现了各种各样有效的机器学习算法。机器学习技术不仅可以对股票价格预测进行海量数据信息的处理，还可以对计算程序进行优化，通过计算机的快速运算，对未来股票价格进行准确预测。

近些年来行为金融学的研究表明， 人类并不具备无限的理性。金融市场上存在大量如过度自信、羊群效应等市场异常现象。将行为、心理因素引入股票预测中可能对预测准确率有影响。投资者的行为表现在他们接受的信息和对待这些信息的态度上， 因此许多学者把目光转向于网络舆情信息（如财务公告、财政新闻、金融研报等） [5-6]，试图探索出它们对股票市场的影响。情感分析技术已广泛应用于股票预测。情感分析的任务是根据输入文本的不同主题， 发现、提取和分类观点或态度。 Boolen 等通过 twitter 对金融数据进行情感分析来研究市场预测， 使用两个心情追踪工具， 通过 6 个心情维度来计算情感值。徐琳等人研究了网络舆情（主要是微博） 对我国股票市场的影响， 提出了网络舆情可能对股票市场具有正效应、负效应和超效应， 并通过实验验证了这三种效应。 Zhai 等结合公司新闻和技术指标， 利用 SVM 分类器对公司股票价格进行预测， 其实用了七个技术指标将公司新闻分为公告新闻和市场新闻两类， 实验结果显示组合的预测效果较好 [7]。

本文基于上述研究成果， 系统地对比了各种机器学习算法和有无情绪数据对特定公司短期股票价格的影响，并选取最优模型对未来股票价格进行预测。

1.2 文章结构

本文将使用 5 种不同的机器学习模型对归票价格进行预测， 并分析各种算法的有效性。根据特定公司（特斯拉公司） 的历史股价数据， 分析算法的准确性， 最终找到并使用最有效的算法来预测股价。同时，还将分析股票买家对股票公司的态度，使预测更加准确。

2 数据和预测方法

2.1 数据

为了进行预测， 我们首先需要收集预测所需的数据。本研究借助 AK-Share 库对特斯拉公司(股票代号 TSLA)1的历史股票数据进行收集。我们收集到了该公司自上市以来每交易日的开盘价、收盘价、最高价、最低价和成交量数据， 并绘制了时间序列图像（**图 1**), 整理了其统计表格（**表 1**）。

1特斯拉（Tesla），是一家美国电动汽车及能源公司， 产销电动汽车、太阳能板、及储能设备。近年来由于电动汽车和自动驾驶领域的卓越表现而受到大众瞩目。

图 1: 特斯拉公司历史股价

表 1: 特斯拉公司历史股票价格信息统计

	open	high	low	close	volume
count	2782.00	2782.00	2782.00	2782.00	2782.00
mean	262.89	268.60	257.07	263.04	8572952.35
std	258.50	266.08	251.10	259.03	12231030.93
min	15.85	16.63	14.98	15.80	114800.00
25%	41.98	42.33	40.59	41.84	2259248.25
50%	227.38	230.84	222.94	227.00	4969275.50
75%	316.73	323.03	311.29	317.28	9056176.25
max	2295.12	2318.49	2186.52	2238.75	172102445.00

本文不仅结合公司的历史股价信息（区别于传统的时序模型如 ARIMA)，同时借助人们在社交网站上的表现出对该公司的情绪数据来增强对未来股票价格的预测作用。首先笔者收集了 2020 年 1 月 1 日值 2020 年 12 月 31 日期间推特上有关特斯拉公司的推文， 并将推文进行自然语言处理（NLP ）之后，转化为情绪打分，并将其结合到股票预测中去。

推文数据是通过爬虫收集“tesla”“TSLA’’等关键词获取的。对于原始的推文数据， 借助 NLTK 库和 Scikit-Learn 去除了其中的停用词、链接和数字等无关信息，并对其绘制了词云图像（图 **2**）, 统计了前 20 个出现频率最大的词语（图 **3**）。

接下来使用 TextBlob 库对清洗后的每条推文数据进行评分（取值范围为 [-1,1], 越接近-1 说明情绪越消极， 越接近于 1 说明情绪越积极）。最终获得对每一天的所有推特用户的情绪打分取平均，获得每天的关于特斯拉公司的平均情绪得分数据。

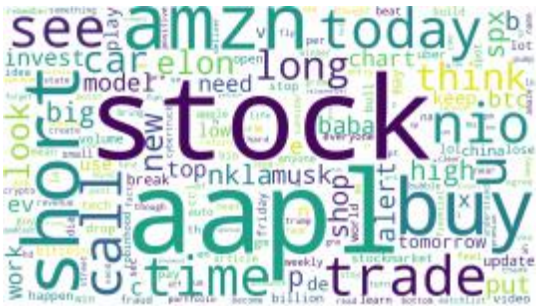

图 2: 推文词云图

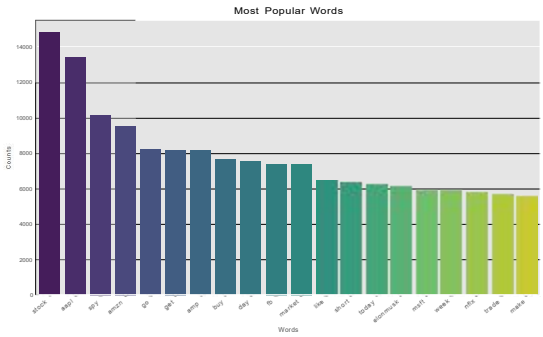

图 3: 前 20 个出现频率最多的

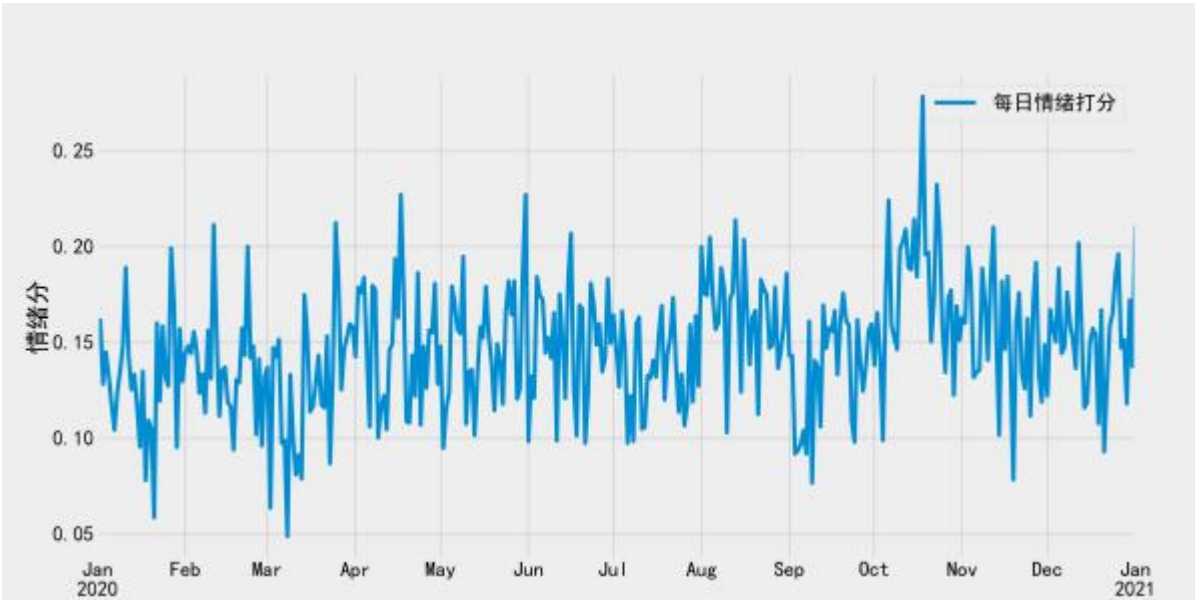

图 4: 2020 年有关特斯拉公司推文的情绪打分

2.2 机器学习模型

2.2.1 K-近邻模型

K-近邻 (KNN) 是一种机器学习算法， 它根据附近的示例做出决策。在统计学中， KNN 算法是一种非参数分类方法。它用于分类和回归。在这两种情况下， 输入都由数据集中最接近的 k 个训练示例组成。输出取决于 k-NN 是用于分类还是回归。一个对象通过其邻居的多数票进行分类和回归， 对象被分配到其 k 个最近邻居中最常见的类或取最近邻居取值的平均值（k 是一个正整数， 通常很小）。 KNN 算法在执行时会经历以下阶段： 1) 将 K 设置为选定的邻居数 2) 计算提供的/测试示例与数据集示例之间的距离 3) 对计算出的距离进行排序 4）获取前 K 个条目的标签或取值 5)返回关于测试示例的预测。

2.2.2 支持向量机

在机器学习中， 支持向量机（SVM) 是具有相关学习算法的监督学习模型， 用于分析数据以进行分类和回归分析。给定一组训练样例， 每个样例都标记为属于两个类别之一， SVM 训练算法构建一个模型， 将新示例分配给一个类别或另一个类别，使其成为非概率二元线性分类器。

SVM 将训练示例映射到空间中的点， 以最大化两个类别之间的差距宽度。然后将新示例映射到相同的空间，并根据它们落在间隙的哪一侧预测属于一个类别。改进的 SVM 算法可用于回归。

2.2.3 决策树模型

决策树模型是运用于分类以及回归的一种树结构。决策树由节点和有向边组成， 一般一棵决策树包含一个根节点、若干内部节点和若干叶节点。决策树的决策过程需要从决策树的根节点开始， 待测数据与决策树中的特征节点进行比较， 并按照比较结果选择选择下一比较分支， 直到叶子节点作为最终的决策结果。

2.2.4 线性回归

线性回归一种用于对标量响应与一个或多个解释变量（也称为因变量和自变量） 之间的关系进行建模的线性方法。一个解释变量的情况称为简单线性回归； 对于不止一种， 该过程称为多元线性回归。在线性回归中， 使用线性预测函数对关系进行建模， 这些函数的未知模型参数是从数据中估计出来的。这种模型称为线性模型。线性回归侧重于给定预测变量值的响应的条件概率分布， 而不是所有这些变量的联合概率分布。线性回归是第一种被严格研究并在实际应用中广泛使用的回归分析类型。

2.2.5 神经网络模型

对于大多数预测而言， 最后一种也是最有效的方法是神经网络或人工神经网络 (ANN)。人工神经网络基于一组称为人工神经元的连接单元或节点， 它们对生物大脑中的神经元进行松散建模。每个连接， 就像生物大脑中的突触一样， 可以向其他神经元传输信号。人工神经元接收信号然后对其进行处理， 并可以向与其相连的神经元发送信号。连接处的“信号”是一个实数，每个神经元的输出由其输入总和的某个非线性函数计算。连接称为边。神经元和边缘的权重通常会随着学习的进行而调整。权重增加或减少连接处的信号强度。神经元可能有一个阈值， 这样只有当聚合信号超过该阈值时才会发送信号。通常， 神经元聚合成层。不同的层可以对其输入执行不同的转换。信号从第一层（输入层） 传输到最后一层（输出层)。本文所采用的神经网络模型是适合进行时序预测的长短期神经网络（LSTM) 模型，其基本结构如下：

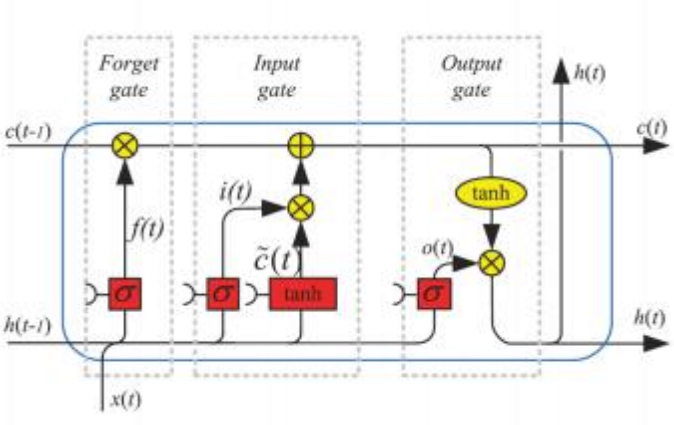

图 5: LSTM 基本结构

2.3 情绪分析

情感分析（也称为意见挖掘或情感 AI）是使用自然语言处理、文本分析、计算语言学和生物识别技术来系统地识别、提取、量化和研究情感状态和主观信息。情感分析广泛应用于客户的声音， 例如评论和调查回复、在线和社交媒体以及医疗保健材料， 应用范围从营销到客户服务再到临床医学。

情感分析的一项基本任务是在文档、句子或特征/方面级别对给定文本的极性进行分类——文档、句子或实体特征/方面中表达的意见是积极的、消极的还是中性的。例如， 高级的“超越极性”情绪分类着眼于情绪状态，例如享受、愤怒、厌恶、悲伤、恐惧和惊讶。

确定情绪的另一种方法是使用标度系统， 其中通常与具有负面、中性或正面情绪相关联的词在 -1 到 +1 范围内（最负面到最正面） 被赋予相关数字或简单地从 0 到正上限， 例如 +0.4。这使得可以调整给定术语相对于在句子层面的情绪。当使用分析一段非结构化文本时， 指定环境中的每个概念都会根据情感词与概念的关联方式及其相关分数给出一个分数。

Python 情感分析是一种分析文本以发现隐藏在其中的情感的方法。它通过结合机器学习和自然语言处理 (NLP) 来实现这一点。基于这一特点， 我们可以对股票评论进行情绪分析， 再求平均值。这可以反应人们当时对于这家公司的整体态度， 以预测股票的变化。将单词转换为其原始形式的方法称为词干提取。当词干提取一个词的语言词根时， 词形还原是将一个词转换为原始形式。在正式进行情感分析之前， 首先进行词干提取， 以便计算每个单词原始形式的出现次数。 TextBlob 是一个用于自然语言处理 (NLP) 的 python 库。本文使用基于词典的方法， 由其语义方向和句子中每个词的强度来定义的“计算”情感。预先定义的字典来分类否定和肯定的词， 根据肯定词语与否定词语出现的次数， 在给所有的单词打分。之后， 通过一些汇集操作（如取所有情绪的平均值）， 计算出最终的情绪得分。 TextBlob 还会对一句话返回句子的极性和主观性。极性在 [-1,1] 之间， -1 表示消极情绪， 1 表示积极情绪。 TextBlob 具有语义标签， 可以帮助进行细粒度分析。例如——表情符号，感叹号，表情符号等等。主观性介于 [0,1] 之间。主观性量化了文本中包含的个人意见和事实信息的数量。较高的主观性意味着文本包含的是个人观点而不是事实信息。 TextBlob 还有一个参数- intensity。TextBlob 通过查看“强度”来计算主观性。强度决定了一个单词是否修饰下一个单词。对于这个特殊的例子， 极性 = -1 和主观性是 1，这是有道理的。

3 实验设计

3.1 实验流程

本研究首先获取了股票数据和舆情数据， 对舆情数据做了又进一步进行了清洗。分别利用上节所列模型对全年前 10 个月的数据（分为含情绪数据和不含情绪数据） 进行训练， 然后在后两个月数据使用所获得的模型，并检验模型的预测效果 (**图 6**)。

3.2 参数设计

本次实验采用的机器学习模型的设置如下所示：

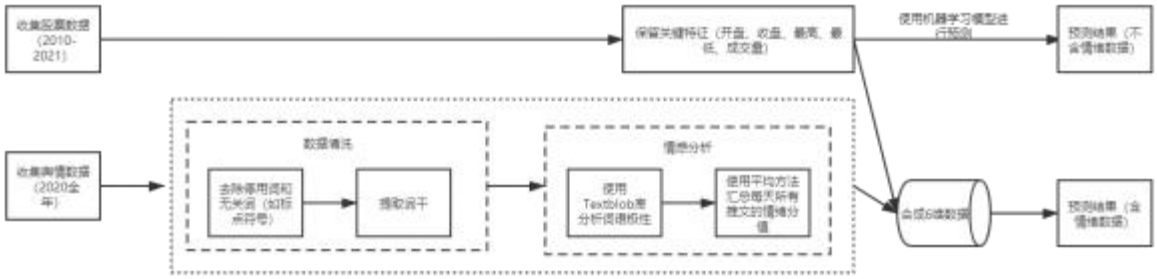

图 6: 实验流程

- 线性回归: fit_intercept =True, normalize =False, copy_X =True,positive =False;
- 支持向量机: kernel = ‘rbf’, degree =3, gamma=’scale’, coef0 =0.0
- K 近邻：n_neighbors=5，weights=’uniform’,algorithm=’auto’，leaf_size=30，metric=’minkowski’
- 决策树： criterion = "mse"，splitter ="best"，max_depth =None，min_samples_split =2
- LSTM:input_size=5(不含情绪数据情况）或=6（含情绪数据情况） ,hidden_size=2,num_layers=1, batch_first=True

4 结果

4.1 准确性

不同的算法对预测同一股票价格有不同程度的准确度。本文使用线性回归、支持向量机、K-近邻、决策树和神经网络（这里用是使长短期神经网络 LSTM ）基于前一天的股票价格信息（开盘价、收盘价、最高价、最低价和成交量） 和情绪打分， 对每个算法进行训练， 并计算出其在预测数据集上的均方误差（MSE)。

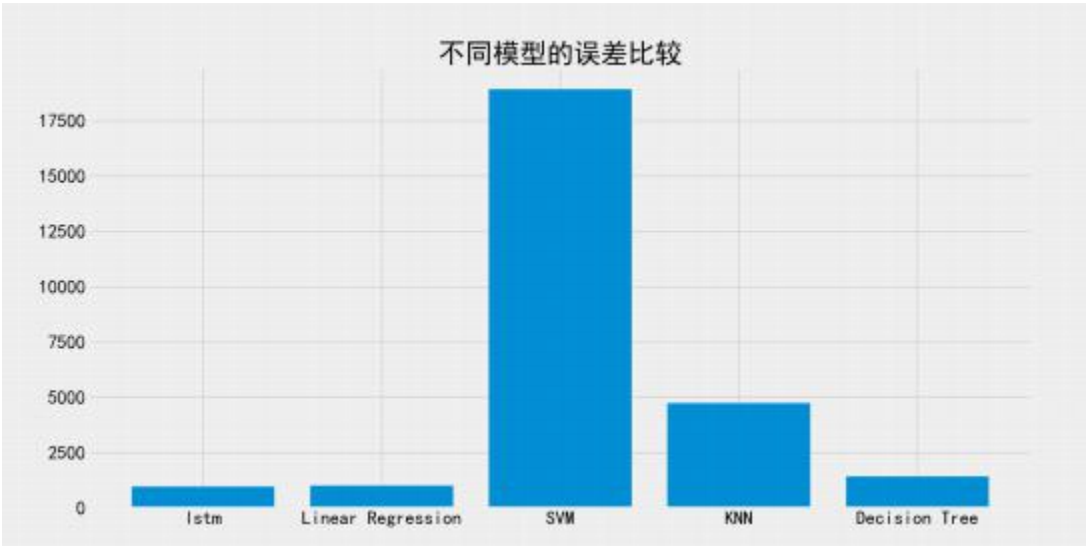

图 7: 不同模型的误差

结果显示， 对于特斯拉股票价格的预测， 准确率从高到低依次为： LSTM, 线性回归， 决策树， K-近邻和支持向量机。

我们以准确度较高的线性回归模型为例， 展示训练模型在全部数据、训练数据和测试数据上的预测效果。

图 8: 线性回归预测

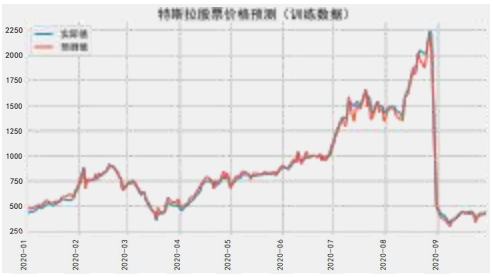

图 9: 训练集预测效果（线性回归）

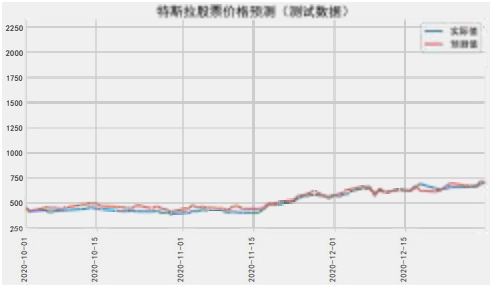

图 10: 测试集预测效果（线性回归）

4.2 有/无情绪

本研究对比了添加情绪数据与不添加情绪数据对股票价格预测之间的误差。结果显示， 添加情绪因素， 在不同模型上效果并不相同。增添情绪数据， 降低了了 LSTM 模型和回归树模型的误差；对于线性回归、支持向量机和 K-近邻模型而言，则增加了在测试集上的预测误差。

5 结论

本文以特斯拉公司 2020 年全年的股票数据和推特用户对特斯拉公司的推文数据为例， 量化了网民对该公司的情绪， 运用多个机器学习模型对未来股票价格进行预测。研究结果表明， 我们可以通过机器学习模型基于历史股价数据对未来股价进行较为准确的预测； 不同机器学习算法对股票价格预测效果不同； 增加情绪数据对不同模型的影响效果不同， 增加情绪数据可以增强深度学习（LSTM) 和线性回归模型的预测准确率， 降低 K 近邻模型、支持向量机模型和决策树模型的准确度。

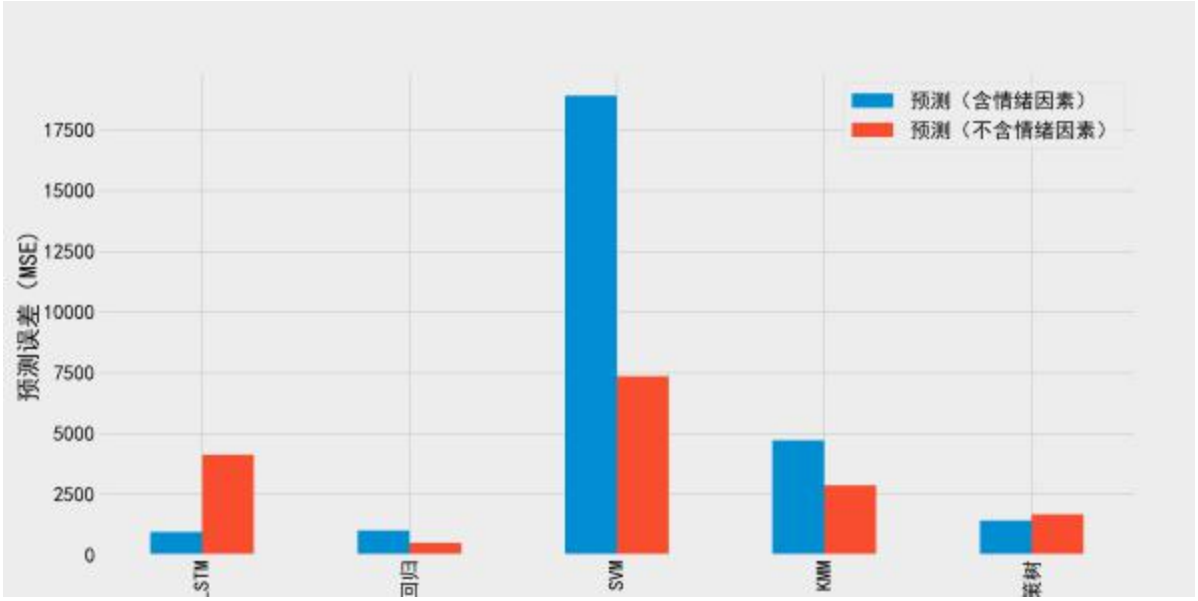

图 11: 比较包含情绪数据与不包含情绪数据的预测误差

关键词: 股价预测, 机器学习, 情感分析, K 近邻, 决策树，支持向量机， LSTM

参考文献

[1] 陈焕, 陈澎. 运用 ARIMA 模型预测股票价格——以莱宝高科 (002106) 为例 [J]. 商情, 2012, 000(008):29-29.

[2] 杨寒冰. 机器学习在股票价格预测中的应用 [D]. 南开大学, 2016.

[3]孙世轩, 潘格格. 基于机器学习的股票特征预测机构持股研究 [J].金融经济,2019,No.518(20):39-43.

[4] 饶东宁, 邓福栋, 蒋志华. 基于多信息源的股价趋势预测 [J]. 计算机科学, 2017, 44(010):193-202

[5]刘震, 王惠敏, 华思瑜, 等. 基于深度学习的股价预测研究 [J].科技创新导报,2018,015(013):247-248.

[6] 邬春学, 赖靖文. 基于 SVM 及股价趋势的股票预测方法研究 [J]. 软件导刊, 2018, 017(004):42-44.

[7] 刘晓东. 沪深 300 股指期货价格预测 [J]. 时代金融, 2017(5):182-183.

[8] 傅魁, 刘玉洁, 陈美丽. 基于财经新闻情感倾向值的股票价格预测 [J]. 北京邮电大学学报 (社会科学版), 2019, 21(01):91-104.

[9] 王新武. 股票价格预测模型 [J]. 陇东学院学报, 2012(03):40-43.