Projekt Del 1

**Student:** Svante Fredin

**Tema för projekt:** Förutsäga börskurs

**Beskrivning av problemet och vad det är du vill åstadkomma:**

Jag vill tjäna pengar genom att köpa en aktie billigt och sälja den när priset gått upp men problemet är att det är svårt att förutse om priset på en aktie kommer gå upp eller ner.

Jag vill lösa problemet genom att vid stängning av dagens handelsdag kunna förutse om priset vid stängning imorgon (eller kommande handelsdag) kommer att vara högre eller lägre. Förväntas priset gå upp så är det en god idé köpa aktien.

Avgränsningar:

* Enbart OMX Stockholm 30 (Yahoo Finance ticker: *^OMX*) kommer att analyseras. Observera att detta är ett index över de trettio mest omsatta aktierna på Stockholmsbörsen och inte en handelsprodukt såsom en aktie eller fond, men jag väljer ändå att analysera det på grund av dess holistiska koppling till handeln på Stockholmsbörsen. Modellen som utvecklas kan sedan användas för analys av enskilda aktier eller fonder.
* Modellen förutsäger om aktien/fonden kommer stiga i värde under kommande handelsdag. Dvs den utgår från att inköp sker vid öppning av kommande handelsdag och försäljning i slutet av samma handelsdag.

**Data:**

Data hämtas genom anrop till Yahoo Finance API och kommer på nedanstående format.

| **Date** | **Open** | **High** | **Low** | **Close** | **Volume** |
| --- | --- | --- | --- | --- | --- |
| 19/02/2019 | 107.79 | 108.66 | 107.78 | 108.17 | 18038460 |
| 20/02/2019 | 107.86 | 107.94 | 106.29 | 107.15 | 21607671 |
| 21/02/2019 | 106.9 | 109.48 | 106.87 | 109.41 | 29063231 |
| 22/02/2019 | 110.05 | 111.2 | 109.82 | 110.97 | 27763218 |
| … | … | … | … | … | … |

* **Är det komplett?**  
  Ja, data är tillräckligt komplett för att kunna bygga en ML-modell av regressionstyp för att förutse börskursen.
* **Har du null-värden?**  
  Ja, samtliga rader med null-värden raderas.
* **Har du extrema värden?**  
  Eventuellt men förmodligen på en försumbar nivå då extremvärden är relativt ovanligt för börskurser. Den modell jag tänkt använda, Random Forest Classifier, kan också hantera en viss nivå av outliers.
* **Vilka datatyper har datat?**  
  String, float, integer.
* **Vilka fält i ditt data vill du använda dig av?**  
  Steg 1:
  + Features: 'Close', 'Volume', 'Open', 'High', 'Low'
  + Target definieras genom att skapa en ny kolumn i vår dataframe som kallas ‘Tomorrow Close’. Med den som input definieras 'Target' = 'Tomorrow Close' > 'Close'

Steg 2:

* + Features omdefinieras till två nya värden med syfte att fånga samband i vår data. Dessa värden beräknas för de senaste 2, 5, 60, 250 och 1000 handelsdagarna:
    - Close ratio:  
      (dataframe[‘Close’] / dataframe.rolling(horizon).mean()[‘Close’])
    - Trend:  
      (dataframe.shift(1).rolling(horizon).sum()[‘Target’])
  + Target behåller samma definition som i steg 1.
* **Hur kan du konvertera alla fält du vill använda till ett numeriskt format?**  
  ‘Target’ blir en boolean men konverteras till en int i form av 0 och 1. Datum används som index och inte som input till modellen, därför behöver inte datumen hanteras.
* **Antal rader och kvalitet på data?**Data håller hög kvalitet och innehåller en rad för varje handelsdygn sen 20 November 2008.

**Typ av problem**

Det är ett regressions-problem med supervised learning. Den modell som kommer användas kallas *Random Forest Classifier* och är en vanligt förekommande modell för att förutse börskurser.

Som tidigare beskrivet tränas två modeller med olika features och deras precision beräknas med biblioteket sklearn’s funktion “precision\_score” som ger andelen korrekta förutsägelser från modellen. Den jämförs med andelen uppgångar som test-datan haft under samma period, vilket kan liknas vid att en person på chans köper aktien varje dag i hopp om att den ska gå upp. Vår förhoppning är att en person som följer vår modell ska få rätt fler gånger än en person som på chans köper aktien varje dag.