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 redan vid öppning av handelsdagen kunna förutse om priset vid stängning kommer att vara högre eller lägre, på så sätt kan jag veta om jag bör köpa aktien vid öppningen och sen sälja den innan stängning.

Avgränsningar:

* Enbart Microsoft Corporation aktie (Nasdaq: MSFT) kommer att analyseras.
* Inköp och försäljning sker under en och samma handelsdag.

**Data:**

Källa: Google Finance, Excel Stocks eller Yahoo Finance på detta 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?**  
  Nej
* **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 Regressor, 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?**  
  Som features använder jag Date och Open. Som target använder jag Close. Att jag inte använder fler features beror på att dom inte är kända förrän i slutet dagen (samtidigt som Close blir tillgänglig).
* **Hur kan du konvertera alla fält du vill använda till ett numeriskt format?**  
  Jag kommer konvertera datum till pandas datetime-format för att hantera dom.
* **Antal rader och kvalitet på data?**Data håller hög kvalitet och innehåller en rad för varje handelsdygn sen början på 1980 talet. Jag överväger att endast använda data från år 2000 och framåt för att minska mängden data som behöver bearbetas.

**Typ av problem**

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